uc3m | Universidad Carlos III de Madrid

Grado Universitario en ingeniería informática

2019-2020

Trabajo Fin de Grado

ESTUDIO SOBRE LA COMPRENSIBILIDAD DE DOCUMENTOS DE PROCEDIMIENTOS ADMINISTRATIVOS EN LA WEB

Javier Del Olmo Ocaña

Tutor/es

Jorge Luis Morato Lara

Lugar y fecha de presentación

7.2.H03

10-03-2020 09:30

Agradecimientos

En primer lugar, agradecer a Jorge Luis Morato por permitirme colaborar en la investigación y desarrollo de este proyecto y sobre todo por la enorme dedicación puesta en este proyecto.

A su vez, agradecer a Vicente Palacios por la implicación en la búsqueda de un proyecto que se ajustase a los requisitos proporcionados para la realización del proyecto.

Por último, pero no menos importante, agradecer a compañeros y familiares la ayuda ofrecida con el análisis de los distintos documentos y valoraciones.

Muchas gracias.

RESUMEN

El fin de es este proyecto es el diseño y la implementación de un sistema capaz de analizar la facilidad de comprensión de textos pertenecientes al ámbito legislativo y de gestiones de documentación en España. Con este sistema, se podrán localizar páginas publicadas por la administración pública las cuales el público al que se enfoca pueda tener dificultades para comprender dichos textos o incluso no cumplir el objetivo el cual es obtener la información que buscan en dicha página.

Durante los últimos años, han ido apareciendo varios portales que ofrecen transcripciones de páginas web estatales con una comprensión más fácil que la publicada por esta. Este fenómeno nos indica que mucha información proporcionada por la administración no se adecua al público objetivo de dicha página.

Estos hechos son los que nos han llevado a desarrollar una herramienta capaz de analizar dichas páginas proporcionadas por el estado y a su vez indicar dependiendo del público objetivo, la facilidad de comprensión del texto.

El objetivo es crear una herramienta la cual permita la valoración de la comprensibilidad de documentos en procedimientos administrativos fue conseguido con éxito mediante el desarrollo de los distintos apartados de la herramienta.

Las capacidades conseguidas para cumplir dicho objetivo son la de poder analizar una página web perteneciente a procedimientos administrativos proporcionada por un usuario, el cual a su vez aporta su edad para poder determinar a qué grupo de audiencia se está enfocando dicho documento y así ofrecer una valoración de la comprensibilidad de dicha página para el grupo de audiencia al que pertenece dicho usuario.

Este proyecto será implementado como un servicio web, el cual contará con una interfaz web para recibir tanto la URL de la página a analizar como de mostrar el indicador de facilidad de comprensión de dicha página a la vez que se proporcionará la edad del usuario.

La página será descompuesta mediante un web Crawler, con el objetivo de obtener las partes de la página útiles para el consecuente análisis. Una vez descompuesta la página,

los datos serán almacenados en una base de datos documental y posteriormente analizados por una inteligencia artificial. El algoritmo ha sido entrenado anteriormente mediante un conjunto de valoraciones de páginas obtenidas.

Tras el análisis de las páginas y la obtención del valor de la comprensibilidad de la página, este será mostrado al usuario por la interfaz web en la que proporcionó la información necesaria.

Los resultados obtenidos han sido prometedores, si bien se pueden implementar mejoras relativas a dominios concretos y subpoblaciones de usuarios.

Con respecto al marco regulador, este queda reflejado en este documento en el apartado con dicho nombre, en el cual se recoge la legislación que aplica al sistema.

ABSTRACT

The scope of this project is to design and develop a system capable of analyzing the ease of understanding of a government website, aimed at electronic administrative procedures. With this system, it may be possible to identify pages that may be difficult for specific users to understand, or that may not even fulfill their objective of guiding citizens on certain procedures.

In recent years, many portals have appeared that offer transcripts of e-government resources, explaining how to interpret the information or use a government web services. The appearance of these portals shows that these resources are not suitable for the target audience

These facts have led to a development of a tool able to analyze those pages provided by the state and depending of the targeted public, the easy of understanding.

The scope of the project is to create a tool that allows the evaluation of the ease in understanding in documents that provide information about administrative processes was reached successfully through the development of the different components of the tool.

The different aspects of the tool that were covered with it to reach the scope of the project were the capability to analyze a web page that belongs to administrative processes of the state provided by a user, which also provides his age to determine the target audience and with this information provide a score of the page depending on the ease in understanding the web page has for this user.

In first place, is needed to define the readability of a text.

A text readability is defined as the ease of understanding a text. This ease of understanding can be provided by the way that the text has been written, the length of the paragraph or phrases, or readers dependent factors.

In this field, there are many tools that generate valuations of the ease in understanding from texts. In the majority of the projects focused to this field, they use linear algorithms that generates the same valuation for the same case without taking changes in the way of writing in count.

Also there are a few projects in this field that are implemented with an Artificial Intelligence but there are less projects that are developed to analyze pages of administrative web processes of the State.

As this field has not been explored for many projects and less in Spanish language and even less with an Artificial Intelligence as a main method of valuations generations, this project will cover some deficiencies of researching in this field.

Also, one of the personal goals acquired with this project was the investigation of different fields.

In first place, the Artificial Intelligence field is a large field with different types of artificial neural networks. The study of the optimal Artificial Intelligence for the system has led to a huge research of the Artificial intelligence that has fulfil the personal goal of investigating in this field.

Regarding the development, the increase of the knowledge of the different languages used in this system has been a great incentive. Especially the freedom of developing something that has been designed by the same person, allowing to change the requisites whenever is needed and using an agile development methodology that has allow a fast development with changes not compromising the time for each phase.

About the database, the researching in NoSQL databases and implementation of these ones was a goal defined with the design of the project. This project was designed to analyze administrative web pages of the State with the optimal technologies for each component which has led to the investigation of different fields. In the databases field, it has led to the NoSQL databases, especially to the documental databases allowing to investigate this kind of databases.

Regarding the initial state of the project, nowadays the readability fields in the texts is being studied in the different languages by project developed by particulars or by institutions, as can be the universities.

Nevertheless, there are few tools that analyzes web pages in Spanish. Also, the majority of tools developed to analyze web pages uses functions or linear methods.

Also there are few researches that are focused on Spanish and even less focused on this kind of documents.

It’s important to remark the methodology implemented for this project. This project has been designed following an agile methodology. The methodology selected for this project is Scrum.

This methodology divides the projects in sprints. Each sprint consists in a lapse of time where a list of tasks has to be completed.

Other important aspect of this methodology are the periodical meetings between the different components of the team. These meetings have been performed between the student in charge of the project and the tutor in charge of the project.

The project has been implemented as a web service, which will be compound by a web interface to receive the URL of the page to be analyzed and to show the value provided from the tool about the easy of understanding of the page. Also, this interface will receive as parameter from the user the age of the user. Once received these parameters, a web crawler will decompound the page obtaining the different important parts of the page useful for the analysis of it.

All the data retrieved from the page and the parameters obtained from the processing of the different parameters, will be stored in a NoSQL documental database.

The artificial intelligence, in this case a neural network, will analyze all the data obtained and stored from the page and will generate an indicator of easy of understanding that the page is.

Once this information is generated will be delivered to the user via the web interface mentioned before.

Referring the scope of this project, the main target to reach is the creation of a tool able to analyze the ease of understanding of pages that belongs to administrative web processes of the State. This will be reached by using Natural Language Processing (NLP) techniques which will allow the system to avoid empty words and be capable to analyze administrative web pages of the State independent of the way they have been written.

These techniques with other techniques of retrieval and accessing to information have been used to reach this target.
One of the main methods was the TFIDF. This measure is compound of two different measures. In the first place, we have the TF, Term Frequency. This measure indicates how much common is a word in the document.

$$tf_{i,j} = f_{i,j}$$

Ecuación 1TF measure [1]

This formula calculates the TF of a term *i* in a document *j* calculating the frequency of itself.

The other measure that compound this method is the IDF, Inverse Document Frequency, that indicates if a term is common in the set of documents or corpus.

$$idf_i = \log (N / n_i)$$

Ecuación 2IDF measure [1]

Being N the number of documents where the term *i* was found and n the total number of documents in the full set of documents.

In practice, both measures are used together by multiplying them, obtaining the TFIDF measure.

$$W_{i,j} = tf_{i,j} \times IDF_i = f_{i,j} \times \log (N / n_i)$$

Ecuación 3 TFIDF measure [1]

These measures are the main formulas used for acquiring the parameters used.

The parameters are calculated successfully obtaining correctly the information provided by the different pages, and applying the mentioned formulas to this information.

Also, an Artificial Intelligence (AI) has been trained with a set of documents previously analyzed by a group of users. The set of documents was compound of pages of the state that belongs to administrative web processes with and arbitrary selection. These users received a set of pages and they return their age and the value of the ease of understanding that these pages have for each user.

This Artificial Intelligence was developed as a perceptron. The perceptron is an artificial neural network compound by three different layers. Input layer, that receives the input parameters, hidden layer, that calculates the different values and the output layer that returns the values calculated.

After the investigation of the Artificial Intelligence, the perceptron was selected due to the efficiency of the artificial neural network, the precision of the approximations generated and the ease of development.

The Artificial Intelligence was developed successfully and generates precise valuations for the administrative web pages of the State. There are some cases that the valuations of the Artificial Intelligence are not enough precise. But the Artificial Intelligence can differentiate between pages with high ease of understanding from the ones that has low ease of understanding.

The data generated by the system and the parameters needed for the generation of the valuations are stored in a NoSQL documental database.

As a documental database, it allows changes in the different documents introduced without any changes in the structure of the database.

As the methodology of the development selected was Scrum, this database has allowed the change of different parameters of the addition of others without any delay in the times proposed for each phase.

The database was successfully developed and works perfectly with the other components because the libraries developed for Python allow to insert, and retrieve data from the database with the structure needed for each step.

The web interface is the one that allows the users to interact with the system. The web interface has two different functions, the first one to receive the request from the users with the parameters needed for the different processes and call to the different system's methods and the second one, to retrieve to the user the results of the different requests.

The interface was developed by different Python libraries successfully and with a short response time.

Regarding the results of the project, the system reaches the goals defined in the scope of the project.

The valuations generated by the system are near to valuations obtained by other systems that calculate the ease of understanding from texts. In some cases, the system provides not much precise valuations. This is due to the lack of documents in the training set of documents, and this is one of the different future improvements detailed in this document.

During the testing phase, it was observed that the parameters that store the age value has a greater weigh on the valuation of the ease in understanding of the page. The same page evaluated with different ages introduced retrieves as readability different values being greater when the age has a greater value.

Also, the parameter that stores the familiarity has a huge impact on the ease in understanding calculation. It was observed that the same page evaluated with different familiarity parameters has a greater value when the familiarity is higher.

Regarding the database, it has successfully covered the requisites that refers to this component, allowing to work with the different documents inserted with a high flexibility of the columns retrieved.

Also, the ease in development of the database connectors with the rest of the components is important to remark.

The efficiency of the database allows to cover the different requisites that refer to the response time of the system. As the system has to retrieve hundreds of documents for the database, is crucial to ensure that the retrieval of these documents allows the system to complete the evaluation of the pages in the estimated time.

The database successfully retrieves the documents in the estimated time to this task and with the structure desired.

Regarding the web interface, it successfully receives the request from the users and retrieve to the user the value of the valuation generated by the system.

The main goal of the web interface is to receive the information provided by the user and show the valuation to the user. This requisite is covered by the system. As the database, in this point is important to take the response time in count.
As the web interface has been developed in the same programming language as the rest of the component are, this made an agile development of the web interface.

Also, the efficiency of the web interface has covered the different requisites that refer to the response time of the system.

An environment and tools used section have been included in this document to describe the environment where the tool was developed and where the system is running. Also all the tools, programming language and libraries have been included in this section describing how they were used and the benefits of using them instead other similar tools or libraries.

An analysis of the system and the initial state of the project have been contemplated. These analyses have been covered in the analysis section which includes, in first place an analysis of the initial state of the project, including different tools and researches performed in this field.

The different use cases have been covered in the use cases section including the different processes that a user can request to the system and the different response the system can provide.

Also in this analysis we can find the different, functional or not functional, requisites designed for this system.

In the test section are defined the different test to be performed in the system, that cover the requisites defined in the analysis section.

The requisites defined in this document have been successfully covered demonstrated in the tests sections.

Also the non-functional requisites as the time response or the portals information to be updated have been successfully covered.

The budget calculated for this project has been contemplated in the planning section. In this section the different phases of the project, and how it was divided in time.

Also the budget with a breakdown of it has been included to this section with the personal costs and the material costs.

Regarding the impact of this project, this project will provide a non-lucrative tool to analyze administrative web pages of the State: The main impact of the system is that this tool will allow to locate administrative web pages of the State that have a lack of readability.

Improving the readability of these page will allow the users of these pages to save time looking for third parties' pages that provides pages with a higher readability.

Also, reducing the amount of money from the users spent on hiring experts in these fields.

To the state, it will accelerate the different administrative processes, receiving the different pays of these processes before allowing to expend these benefits in other projects.

The regulatory framework is contemplated in this document in the section with this name. In this section, all the laws and regulations applied to this system are collected.

During the development of the system, have appeared many implementations that will improve the different components of the system or adding new functionalities to it. All these improvements have been added to this document in the future works section.

In the end of this document there are some sections referring the glossary, the acronyms and a section destined to the bibliography.

ÍNDICE

ÍNDICE DE ILUSTRACIONES

ÍNDICE DE TABLAS

ÍNDICE DE ECUACIONES

INTRODUCCION

La principal motivación de este proyecto es la dificultad de comprensión de las páginas del Estado, que viene mostrado por el crecimiento de número de portales que facilitan la comprensión del texto perteneciente a información del Estado. Esto es un indicador de la ineficacia de dichas páginas ya que el público objetivo no consigue obtener la información necesaria y busca otras fuentes donde poder obtener dicha información. El objetivo es poder localizar dichas páginas que no son comprensibles para ciertos grupos de la población y así poder ser estudiados a la vez que mejorarlos para conseguir el objetivo de dichas páginas, el cual es informar al ciudadano.

A su vez, aunque esta herramienta está enfocada a páginas del estado, se podría generalizar para poder ser útil para herramientas de cálculo de posicionamiento web, calculándose la facilidad de comprensión de un texto y posicionando dicha página en una posición más elevada como se ve reflejado en el artículo de D.Bilal y L.Huang [1].

El entendimiento de la capacidad comprensiva de cada uno de los grupos sociales existentes hoy en día es un indicador de la cultura de una sociedad, por lo que analizando cada uno de los textos en función del entrenamiento proporcionado a la red neuronal, es posible obtener información sobre la población y su nivel de formación, en función de la edad.

El impacto de este proyecto abarca desde la mejora de las páginas gubernamentales hasta la investigación de las limitaciones de la comprensión por los distintos grupos sociales. Estas mejoras impactarán en la optimización de la información al usuario y en la implantación de la administración electrónica.

Hoy en día se encuentran distintos servicios web que te permiten obtener una puntuación sobre una página web en concreto, permitiéndote ver aspectos como legibilidad o si cumple ciertos estándares definidos. El proyecto surge a raíz de que no existe ninguna página en español capaz de realizar una medición de la facilidad de comprensión de textos en ámbitos específicos, en este caso existen recursos similares como pueden ser legible.es el cual es un recuso genérico, es decir no es específico del campo de las páginas web del

estado. A su vez, podemos encontrar páginas como lecturafacil.net que generan recomendaciones sobre cómo se pueden mejorar la comprensibilidad de una página web. Existen diversos proyectos de habla inglesa que abordan este problema, pero al no utilizar el español, en este aspecto no hay ningún servicio que cubra esta necesidad.

ESTADO DEL ARTE

En este apartado se describe el estado actual del entorno que rodea los distintos ámbitos que afectan al sistema desarrollado y las distintas opciones posibles que existen en el momento. A su vez, se definen las herramientas definidas para el cálculo de la comprensibilidad de textos en otras herramientas.

COMPRENSIBILIDAD DE UN TEXTO

En primer lugar, para poder estudiar la comprensibilidad de un texto, es necesario definir el concepto de comprensibilidad de un texto.

La comprensibilidad de un texto se define como la facilidad de entender un texto. Esta facilidad puede ser proporcionada por la forma en la que se ha escrito, el tamaño de los párrafos o frases a la vez que, de factores dependientes del lector, según el artículo de la Universidad del País Vasco. [1]

El término comprensibilidad se traduce al inglés como *readability*. Frecuentemente se traduce *readability* al español como comprensibilidad de un texto o legibilidad lingüística. En realidad, *readability* se corresponde con comprensión de texto o lecturabilidad. El término inglés *legibility*, traducido al español como legibilidad, abarca las características de organización y formato del texto (tamaño fuente, contraste con el fondo, etc.), pero no su comprensión.

METODOLOGÍA PARA EL CÁLCULO DE LA COMPRENSIBILIDAD EN OTRAS HERRAMIENTAS.

Una de los principales métodos de cálculo de la facilidad de comprensión de un texto es el Flesch Reading Ease y el Flesch-Kincaid Grade. En 1948, Rudolph Flesch, consultor

de Associated Press desarrolló métodos para mejorar la facilidad de comprensión de los periódicos como se comenta en la página readable.com [2].

Dichos métodos se siguieron utilizando los años siguientes. Este test, devuelve un valor entre 1 y 100 indicando la facilidad de comprensión que tendría un texto.

Este método fue desarrollado por la US Navy creándose así la fórmula Flesch-Kincaid Grade. El principal uso de dicha fórmula fue para la medición de la dificultad de comprensión de documentos de la US Navy que utilizaban para sus entrenamientos.

Flesch Reading Ease

$$206.835 - 1.015\left(\frac{\text{total words}}{\text{total sentences}}\right) - 84.6\left(\frac{\text{total syllables}}{\text{total words}}\right)$$

Flesch-Kincaid Grade Level

$$0.39\left(\frac{\text{total words}}{\text{total sentences}}\right) + 11.8\left(\frac{\text{total syllables}}{\text{total words}}\right) - 15.59$$

Ilustración 1 Flesch fórmulas [3]

HERRAMIENTAS SOBRE LA COMPRENSIBILIDAD DE LOS TEXTOS

Estas dos fórmulas son la base que utiliza el portal web *legible.es* [3], junto con otras fórmulas para su adaptación al español. Dependiendo del tipo de texto que se desee analizar, se utilizarán unas u otras, perteneciendo la mayoría al ámbito escolar, sin embargo existen excepciones como INFLESZ de Inés Barrio [5] que la comprensibilidad de los textos se calibró mediante documentos médicos. Esta herramienta asegura que en el caso de Flesch-Kinkaid, si un texto obtiene una puntuación de 80, el 80% de los americanos será capaz de poder leer dicho texto sin dificultad.

A continuación, se muestran los distintos umbrales de la adaptación de Fernández Huerta del método Flesch:

Puntuación	Dificultad	Estudios mínimos capacitados
90-100	muy fácil	4º grado
80-90	fácil	5º grado
70-80	algo fácil	6º grado
60-70	normal (para adulto)	7º u 8º grado
50-60	algo difícil	preuniversitario
30-50	difícil	cursos selectivos
0-30	muy difícil	universitario (especialización)

Tabla 1Umbrales Fernández Huerta

La herramienta legible.es está diseñada para textos en español, por lo que cubre ciertos aspectos de los cuales queremos aproximar con dicha herramienta. A su vez, utiliza algoritmos ya creados y muy simples, por lo que, si en algún momento dichos algoritmos se invalidaran, la herramienta quedaría desechada. Al contrario que al utilizar una inteligencia artificial a la hora del análisis de los documentos en el que, si por una razón la corriente de aprendizaje o de escritura de los textos cambiase, esta herramienta no sería capaz de poder adaptarse a los cambios.

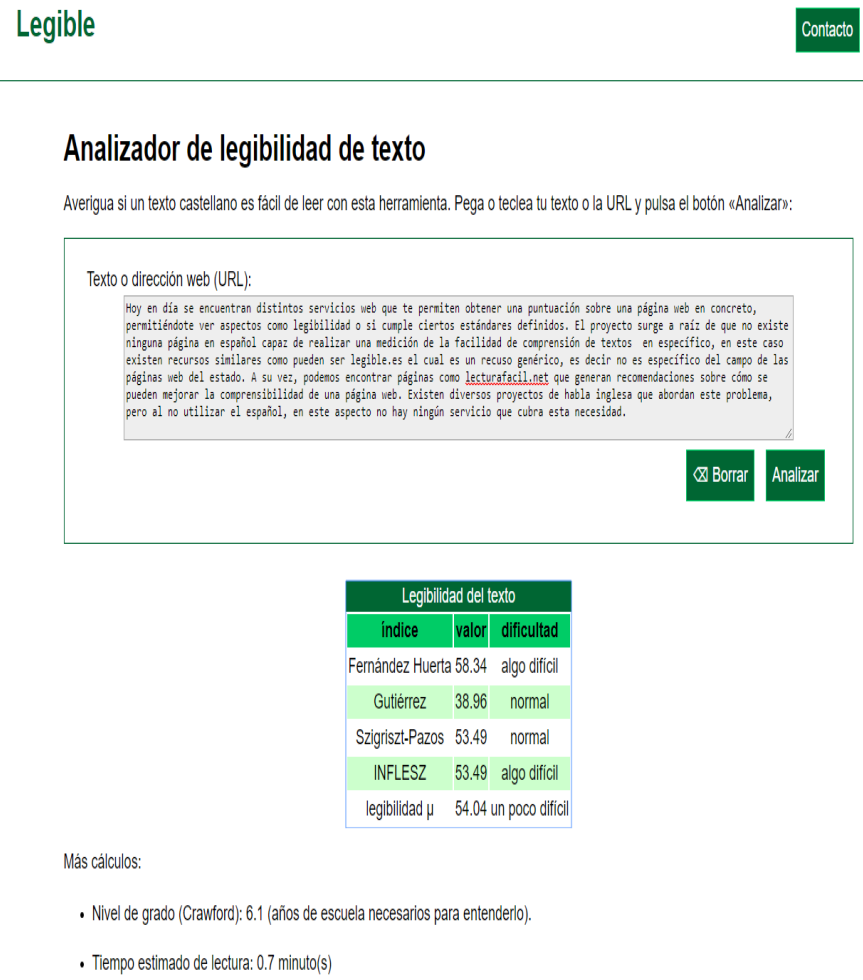


Ilustración 2calculo de legibilidad legible.es [5]

Como podemos ver, es una herramienta de fácil uso, la cual permite al usuario de manera sencilla el análisis de un texto mediante copiar el fragmento de texto que se quiere analizar en el recuadro reservado para ello y pulsar el botón analizar.

Como se puede observar, este servicio web devuelve el cálculo de la comprensibilidad para diversos métodos de cálculo de la comprensibilidad de los textos.

A su vez, se puede observar que, para un texto pequeño, la respuesta de la página es muy buena y tras múltiples pruebas con diversos textos, se ha comprobado que la herramienta en general muestra un tiempo de respuesta aceptable para todo tipo de textos independientemente de su longitud.

Estadísticas del texto	
caracteres	850
letras	702
sílabas	296
palabras	138
frases	6
párrafos	1
letras por palabra	5.09
sílabas por palabra	2.14
palabras por frase	19.71

Palabras en orden de frecuencia		
orden	palabra	frecuencia
1	de	8
2	en	6
3	que	5
4	web	4
5	una	4
6	no	4
7	es	4
8	página	3
9	como	3
10	el	3
11	este	3
12	se	2
13	sobre	2
14	español	2
15	la	2
16	especifico	2

Ilustración 3 Calculo de legible.es resultados [5]

Esta herramienta devuelve diversos parámetros con los que calcula la comprensibilidad de la página, a la vez que pueden ser útiles para el usuario.

Palabras raras o mal escritas	
orden	palabras
1	legibilidad
2	lecturafacil
3	net
4	comprensibilidad

Número de palabras por su número de sílabas	
Número de sílabas	Número de palabras
1	56
2	30
3	37
5	5
4	8
6	2

Letras ordenadas por frecuencia		
orden	letras	frecuencia
1	e	116
2	a	61
3	o	57
4	n	57
5	s	55
6	i	54
7	r	40
8	c	39
9	d	33

Ilustración 4 Cálculo de legible.es resultados 2 [5]

Como se comentaba anteriormente en la definición de comprensibilidad de un texto, el número de palabras por párrafo, por frase o el número de silabas, es un indicador de la comprensibilidad de un texto el cual es mostrado por dicha página a la vez que es usado en diversos métodos para el cálculo de la comprensibilidad de dicho texto.

La herramienta G-Rubic [4] es un portal diseñado por la UNED, capaz de calcular tanto la legibilidad como la facilidad de comprensión de un texto. Está diseñada principalmente para alumnos y estudiantes en general. Lo que permite esta herramienta es analizar textos evaluándolos y devolviendo a la vez que una puntuación de la legibilidad del texto y de la comprensión, comentarios sobre el propio texto que podrían mejorar el mismo en estos aspectos. Esta herramienta se basa en modelos basados en reglas sobre todo centrados en el estudio del procesamiento del lenguaje natural.

Esta herramienta cubre una serie de aspectos que buscamos afrontar con este proyecto, sin embargo, el proyecto que se presenta en este documento busca la calificación de páginas web completas no solo el texto de la misma. A su vez, al igual que en la herramienta anterior, el uso de modelos basados en reglas limita la evolución de dicha herramienta a la hora de adaptarse a cambios en el lenguaje.

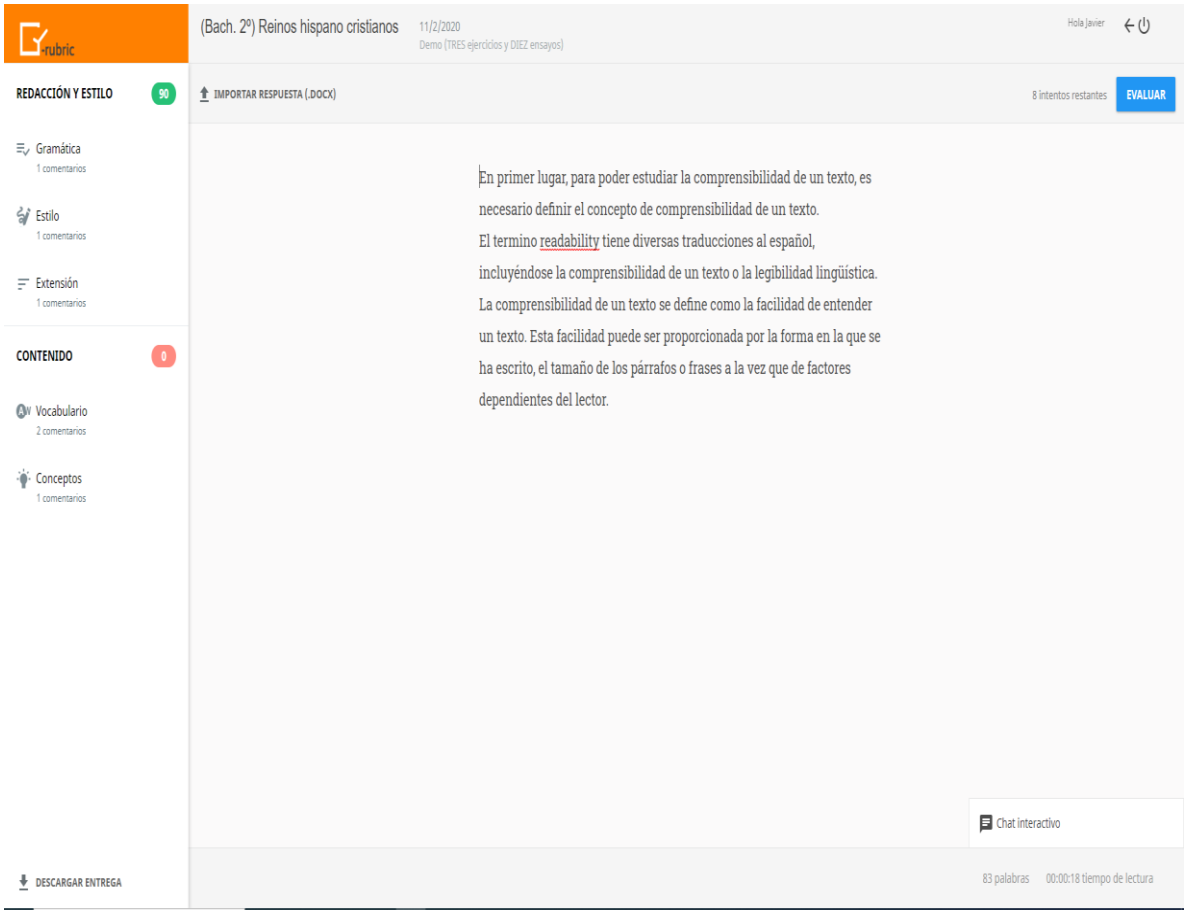


Ilustración 5 G-Rubic interfaz 1 [7]

Como puede observarse en la imagen, la interfaz de G-Rubic es una interfaz muy simple en la cual introducimos el texto a evaluar, y con pulsar el botón evaluar, nos permite obtener una evaluación del texto insertado.

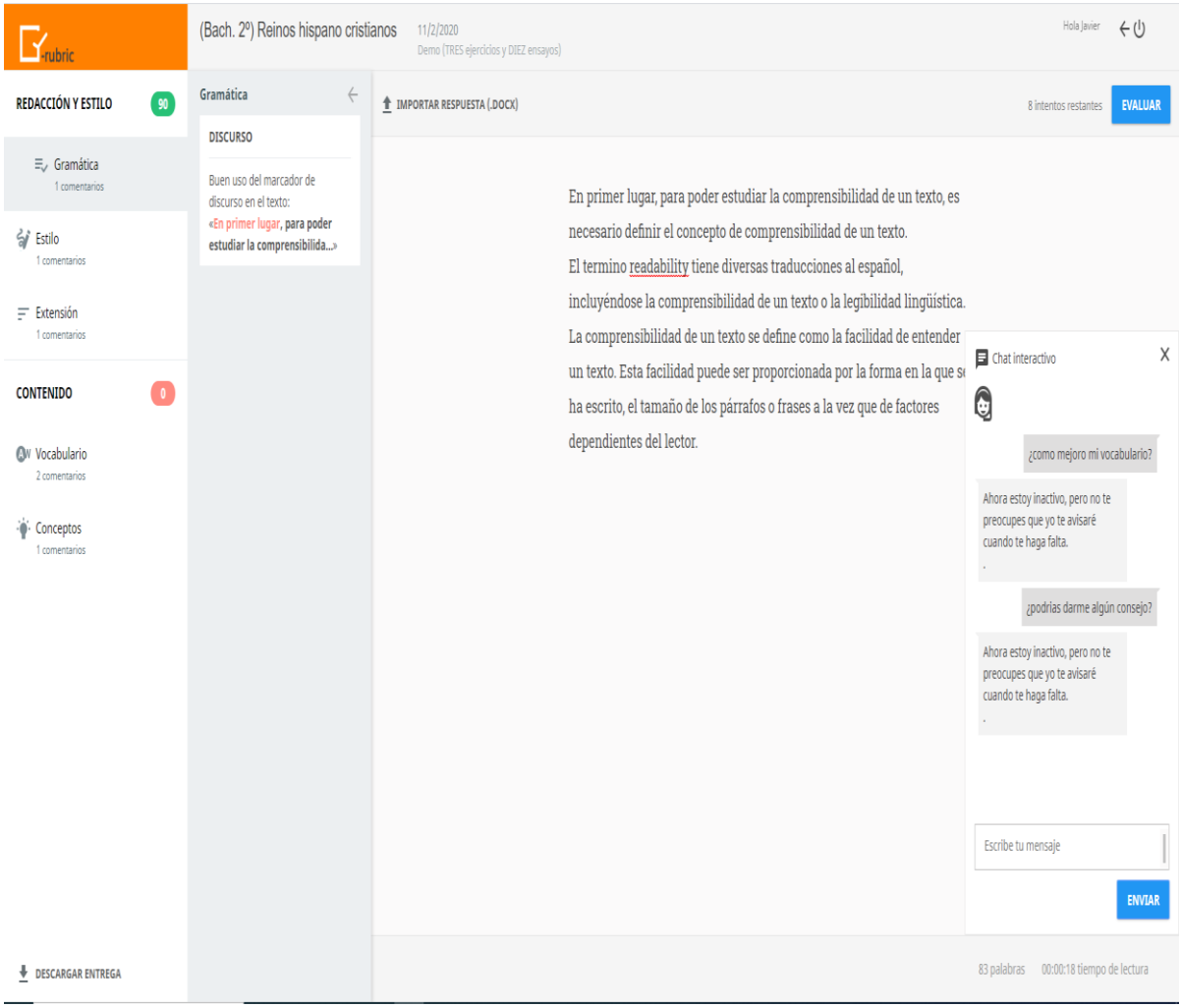


Ilustración 6 G-Rubic resultados [7]

A su vez, la evaluación se realiza mediante diversos aspectos. En primer lugar, nos encontramos con una evaluación de la redacción y estilo en la cual recibimos una puntuación por ella (en el caso de la imagen se ha recibido un 90 en redacción y estilo). Como se puede observar, se divide cada una de las dos valoraciones, en este caso redacción y estilo y contenido, en varios subgrupos en los cuales se ofrecen comentarios sobre los errores o aciertos que hemos podido cometer durante el desarrollo del texto.

A su vez, en la esquina inferior derecha de la imagen podemos observar un chat, que desgraciadamente, durante todas las pruebas que se han realizado en esta página no estaba operativo, en el cual es posible entablar una conversación con un *bot* el cual contestará a cualquier comentario o petición que se le solicite acerca del texto.

Esta herramienta aún está en desarrollo, por lo que con respecto a contenido no es posible analizar cualquier texto, simplemente un conjunto de temas preseleccionados por la herramienta.

Otra de las herramientas del sector con mucho potencial y realmente interesante es TRAY readability tool. Esta herramienta se ejecuta directamente en el navegador de *Chrome* y se instala como una extensión del mismo en el artículo de Robert Korntheuer [5].

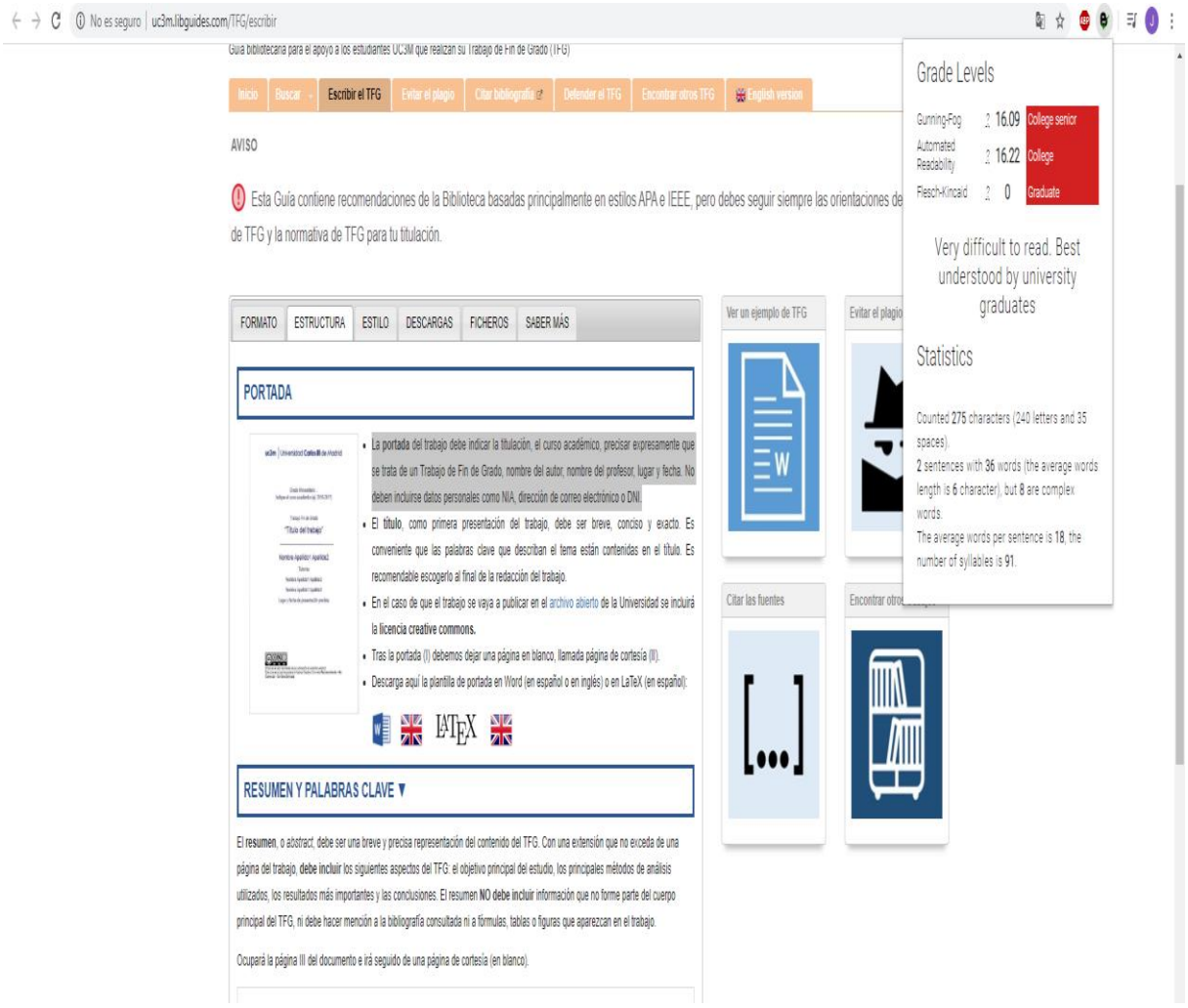


Ilustración 7 Tray readability tool

Esta herramienta solo permite el análisis de textos en inglés, sin embargo, ha resultado interesante introducirla en el estado del arte ya que la facilidad con la que se puede realizar evaluaciones de la comprensibilidad de un texto, la hace destacar entre el resto de herramientas.

Como se puede observar en la imagen, la herramienta aporta valoraciones para tres grupos de público distintos, junto con varios parámetros utilizados en la medición a la vez que pueden ser útiles para el lector.

OBJETIVOS

En este apartado se define el objetivo de este proyecto, indicando la finalidad del mismo y que motivaciones han llevado a querer desarrollar dicho proyecto.

A su vez, se incluye un apartado con los objetivos personales que han llevado a tomar las distintas decisiones sobre la realización de este proyecto.

OBJETIVOS DEL PROYECTO

El objetivo principal de este proyecto es la creación de una herramienta capaz de analizar páginas pertenecientes a procesos burocráticos del estado con el fin de poder obtener una valoración de la comprensibilidad de dichas páginas en función de varios parámetros definidos a lo largo de este documento.

De este objetivo principal, derivan distintos objetivos entre los cuales se comprende el partir de un conjunto de páginas web previamente analizadas, valorar páginas no incluidas en este conjunto de páginas web.

A su vez, existen herramientas capaces de generar análisis sobre páginas web, pero utilizando fórmulas lineares que no permiten un aprendizaje automático por parte de la herramienta.

Un ejemplo de este tipo de fórmulas lineares es el índice de SMOG como referencian en la página *readable.com* [8]. Esta ecuación indica el número de años cursados de media necesarios para poder entender un texto.

$$\text{Nivel de educacion} = \sqrt{\text{Palabras Polisilabas X } \frac{30}{\text{Numero de sentencias}}} + 3{,}1291$$

Ecuación 4Índice de Smog

Como se puede observar, la formula siempre el mismo resultado para un mismo texto.

Por lo tanto, al introducir una inteligencia artificial como medio de análisis, se consigue obtener una herramienta con una evolución constante, capaz de obtener nuevos parámetros para la evaluación de los textos.

Otro de los objetivos, derivados del principal, es una mejora de la generación de textos burocráticos, identificando los que tienen una menor facilidad de comprensión. De esta manera se facilitará el trabajo de ciudadanos los cuales son los principales afectados de la poca facilidad de comprensibilidad de los proporcionados por el estado.

OBJETIVO PERSONAL

Como objetivo personal, la principal motivación de dicho proyecto es la investigación de las distintas técnicas de obtención de información junto con su consecuente procesamiento del lenguaje natural. A esto se suma la investigación de las distintas bases de datos y cómo puede afectar el eficiente almacenamiento de los archivos a la efectividad de la herramienta. Por último, destacar como objetivo la investigación de las diferentes inteligencias artificiales y la evolución de las mismas.

Principalmente, el proyecto en cuestión abarcaba todos los campos estudiados durante la carrera que han sido interesantes para mi persona, por lo que el poder desarrollar una herramienta que me permitiese abarcar estos campos ha sido la mayor motivación por elegir este proyecto.

METODOS

En este apartado se definirá como se pretende conseguir desarrollar el sistema y los distintos pasos que se han dado para llegar al desarrollo final.

METODOLOGÍA IMPLEMENTADA

La metodología aplicada para el desarrollo de este sistema ha sido Scrum. Esta metodología se ha elegido debido a la eficacia del proceso iterativo que para el desarrollo de esta herramienta se acoplaba a la perfección al proyecto ya que permite un desarrollo más ágil, a la vez que permite cambios en el desarrollo del mismo en función de resultados obtenidos en cada una de las distintas fases del desarrollo, como comenta Mario Araque en su artículo [6].

En este caso, las distintas reuniones del equipo se han realizado entre el alumno y el tutor de dicho proceso en el cual se ha definido el orden de desarrollo de los componentes y las distintas mejoras del sistema en función de distintos aspectos como la importancia de los requisitos asociados a dichas funciones o el coste en tiempo del desarrollo de las funciones de cada mejora.

La característica principal de la metodología de Scrum son los *sprints*. Un *sprint* es un periodo corto de tiempo, no suele ser mayor a un mes en el cual el equipo de desarrollo se compromete a entregar un producto que cumpla una serie de requisitos ofrecidos por el cliente.

En la metodología ágil Scrum, existen tres roles los cuales son, propietario del producto, scrum master y equipo de desarrollo [7].

En este caso los tres roles han sido llevados por una misma persona, diferenciando las tareas de cada uno de los roles y ejecutándolas en el orden necesario.

En primer lugar, se han obtenido los requisitos de la herramienta, los cuales son proporcionados por el propietario del producto y recibidor por el equipo de desarrollo y el Scrum master.

La obtención de requisitos es un proceso iterativo en el cual una vez proporcionados los requisitos por el propietario del producto el equipo de desarrollo se encarga de decidir en dicho sprint a que requisitos se comprometen a cumplir.

Estos requisitos obtenidos serán traducidos en tareas para los desarrolladores. Estas tareas serán lo más atómicas posibles, es decir, se tratará de que las tareas sean lo más específicas posibles, evitando cualquier ambigüedad con el fin de poder determinar el esfuerzo que requerirá cada una de ellas con la mayor precisión posible.

El hecho de poder precisar el esfuerzo para cada una de las distintas tareas que corresponden a los distintos requisitos presentados por el propietario del producto facilita el hecho de poder determinar los requisitos que podrán ser cumplidos por el equipo de desarrollo en cada sprint, y no comprometerse a realizar tareas o cumplir requisitos que no es posible cumplir.

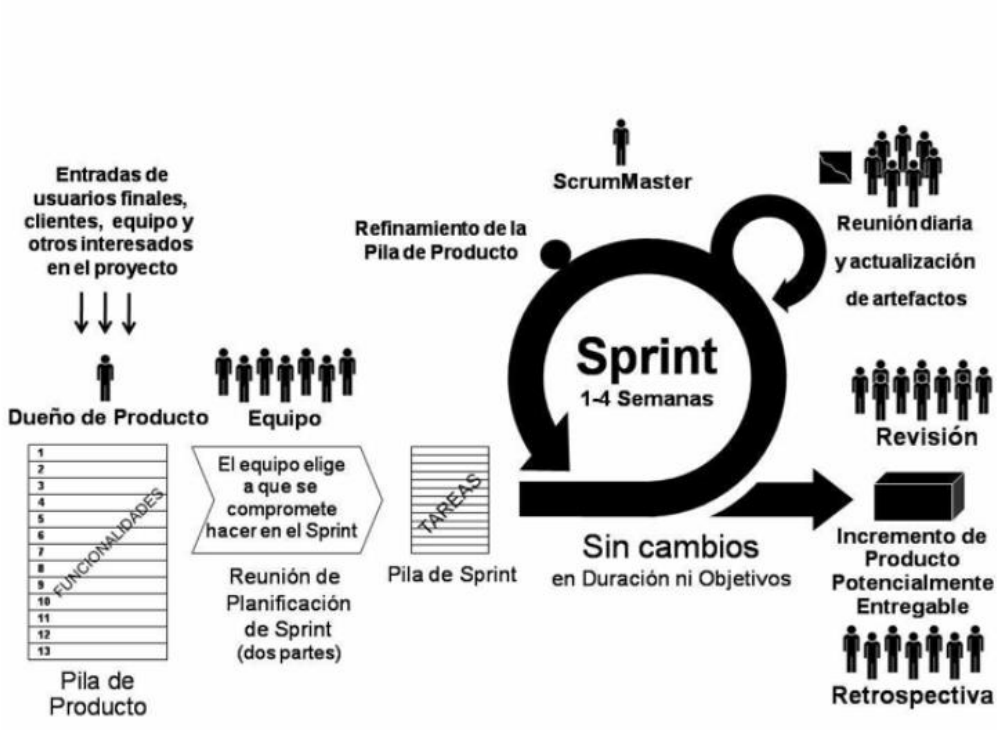


Ilustración 8Metodología scrum [6]

Una vez definidas cada una de las tareas que se desarrollarán en dicho sprint se procede al desarrollo de las tareas definidas almacenadas todas ellas en un *backlog* de producto.

Una vez terminado el sprint, el propietario del productor se encargará de revisar el resultado obtenido de las tareas, dando su opinión sobre el producto y comprobando que requisitos han sido cumplidos y cuales han de ser cambiados en función de los resultados obtenidos.

Este proceso resulta óptimo para el desarrollo de dicha herramienta, sobre todo a la hora de la priorización de tareas, ya que, en primer lugar, al trabajar con una inteligencia artificial, la precisión de las valoraciones generadas por la misma depende de muchos aspectos, por ejemplo, los parámetros introducidos o el entrenamiento de la misma.

Al terminar un sprint y comprobar que los parámetros introducidos a la inteligencia artificial no han afectado a su precisión o la precisión no es la suficiente, es posible cambiar estos requisitos con el fin de tener un producto final acorde a las peticiones del propietario del producto.

ALCANCE DE LA HERRAMIENTA

En primer lugar, se ha de definir la facilidad de comprensión de un texto. La comprensión de un texto es totalmente subjetiva, es decir, depende completamente de la persona, su formación o su ámbito, por lo que no hay una medida estándar de la comprensibilidad de un texto, sin embargo, si se han encontrado una serie de patrones a la vez que se han diseñado otros que nos permiten medir la comprensibilidad de un texto.

La primera fuente de datos para la medición son los propios portales que muestran páginas del Estado, facilitando la comprensión de dichos textos mediante la simplificación o aclaración de los mismos. Si existe una página gubernamental simplificada en uno de estos portales, será un indicador de que al menos un grupo de personas no es capaz de

obtener la información necesaria del texto expuesto por el estado en su página web y se ha visto la necesidad de recurrir a un portal distinto para obtener dicha información.

Otro de los aspectos que toma importancia en dicho cálculo es el uso de tecnicismos. En el caso de una persona formada en un campo específico, la compresión del texto aumenta notablemente debido a la familiaridad con dichos términos, sin embargo, para el resto de personas que no tienen dicha formación esto puede crear incomodidad e incluso incomprensión de dichos términos, haciendo que la comprensibilidad del texto sea nula.

Otro de los valores tenidos en cuenta para el cálculo de la facilidad de comprensión es la familiaridad de las palabras utilizadas en el texto. Se denomina familiaridad al contacto habitual o profundo conocimiento que se tiene de algo. Cuando nos referimos a la familiaridad en el contexto del lenguaje, nos referimos a lo habitual que es una palabra o frase y al conocimiento que tenemos de la misma. Cuando más familiares o frecuentes en el lenguaje sean las palabras en un texto, más fácil de comprender será el mismo.

La cantidad de palabras en una frase o en un párrafo afecta a su vez a la comprensibilidad de un texto. Diversos estudios confirman que textos con frases con menos palabras son más fáciles de entender [8].

Estos parámetros obtenidos de una página, se almacenarán en una base de datos documental.

Al ser una base de datos documental, se facilitará el almacenamiento de gran cantidad de datos para su posterior análisis. En esta base de datos existirá una sola base de datos, en la cual encontraremos cada uno de los documentos analizados. Cada documento contendrá cada una de las palabras contenidas en el texto, al igual que si dicho documento contenía imágenes y en número de ellas junto con los enlaces encontrados en la misma [9].

Uno de los parámetros que se calcula tras la obtención de los parámetros es TFIDF [10]. TFIDF es un indicador formado por dos indicadores a su vez.

En primer lugar, nos encontramos el indicador TF, el cual se traduce del inglés como frecuencia de un término.

$$tf_{i,j} = f_{i,j}$$

Ecuación 5TF [14]

Esta fórmula nos indica la frecuencia del término *i* en el documento *j*.

A su vez, se encuentra el indicador IDF, el cual se traduce del inglés como frecuencia inversa en los documentos.

$$idf_i = \log (N / n_i)$$

Ecuación 6IDF [14]

Siendo N el número de documentos en los que aparece el termino *i* y n el número total de documentos en el corpus. Este indicador nos indica lo frecuente que es un término en el resto de la colección.

Estos indicadores en la práctica se utilizan en conjunto multiplicándose, dando lugar al indicador TFIDF.

$$W_{i,j} = tf_{i,j} \times IDF_i = f_{i,j} \times \log (N / n_i)$$

Ecuación 7TFIDF [14]

En la base de datos encontraremos dos colecciones distintas, por un lado, la colección de páginas valoradas, la cual contendrá todas las páginas ya valoradas, tanto por la inteligencia artificial, como las páginas valoradas por terceros las cuales se han utilizado como entrenamiento de dicha inteligencia artificial. Junto con los parámetros ya comentados en esta base de datos se ha almacenado a su vez el número de palabras por párrafo, el número de palabras por frase y el TFIDF de la página a analizar con las páginas con un título similar al de la página web a analizar encontradas en varios portales que se dedican a proporcionar transcripciones de páginas del estado con una mayor facilidad de compresión.

Una vez almacenados todos los datos, se procederá al cálculo de los anteriores parámetros. En primer lugar, se obtiene el título de la página a analizar y mediante un web crawler se obtienen todos los datos de la página web. Se descarga la página web en versión HTML. A continuación, se procede a guardar los valores importantes, como por ejemplo el título, en la base de datos. A su vez, se realiza una limpieza de todo el código HTML que no aporta valor para la obtención de parámetros para su posterior procesamiento. Una vez obtenidos el texto de la página web, se procederá a calcular el número de palabras por frase y por párrafo. De este número se calcula la media de palabras por frase y por párrafo y se almacena en la base de datos como parámetro.

A su vez, con el título de la página web a analizar, se utiliza la misma como consulta y se procede a calcular el TFIDF del título de la página web, frente a cada una de las páginas web de los distintos portales.

En este punto es importante definir que es un web crawler, según el artículo de Marc Najork [11], un web crawler es un programa el cual permite la descarga de páginas web a la vez que se analizan los distintos elementos de dichas páginas web y se obtienen los enlaces web a otras páginas web. A su vez, estas páginas web se analizarán y se tratarán, produciéndose así un proceso recursivo.

Para la obtención de las distintas páginas web de los portales, previamente se utilizará un spider creado mediante Scrappy [12]el cual se encargará de, dada una URL inicial de cada uno de los portales a analizar, en este caso son los siguientes, junto con un ejemplo de cada uno de estos portales y su correspondiente página perteneciente al Estado:

- Rankia.com

Ilustración 9Rankia.com declaración renta [17]

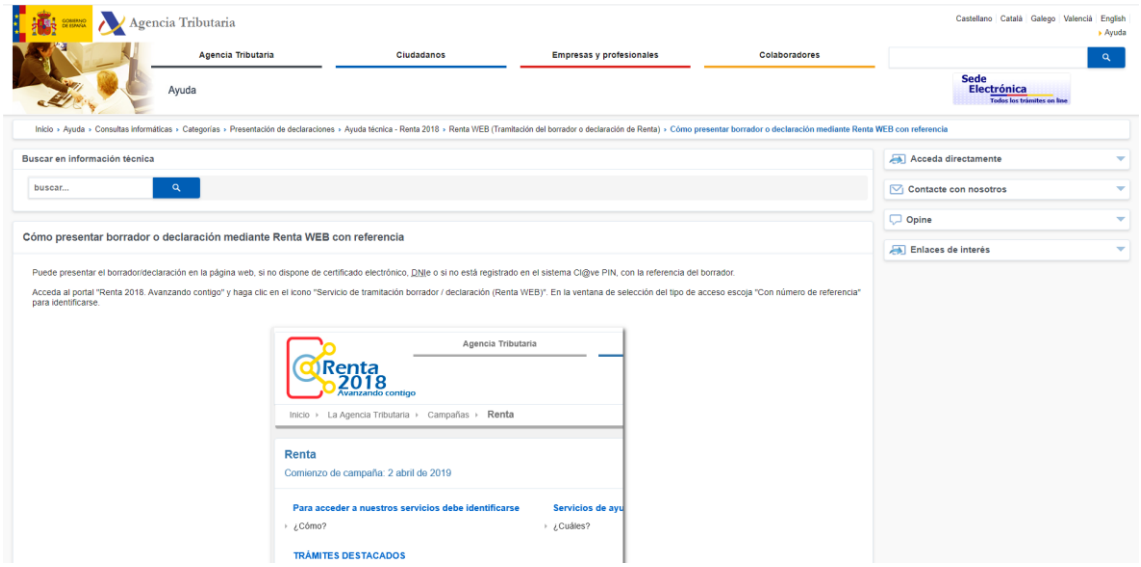


Ilustración 10Agencia tributaria Renta [18]

- Comosetramita.com

Ilustración 11Comosetramita.com Libro familia [19]

Ilustración 12Exterior.gob libro familia [20]

- Adminfacil.es

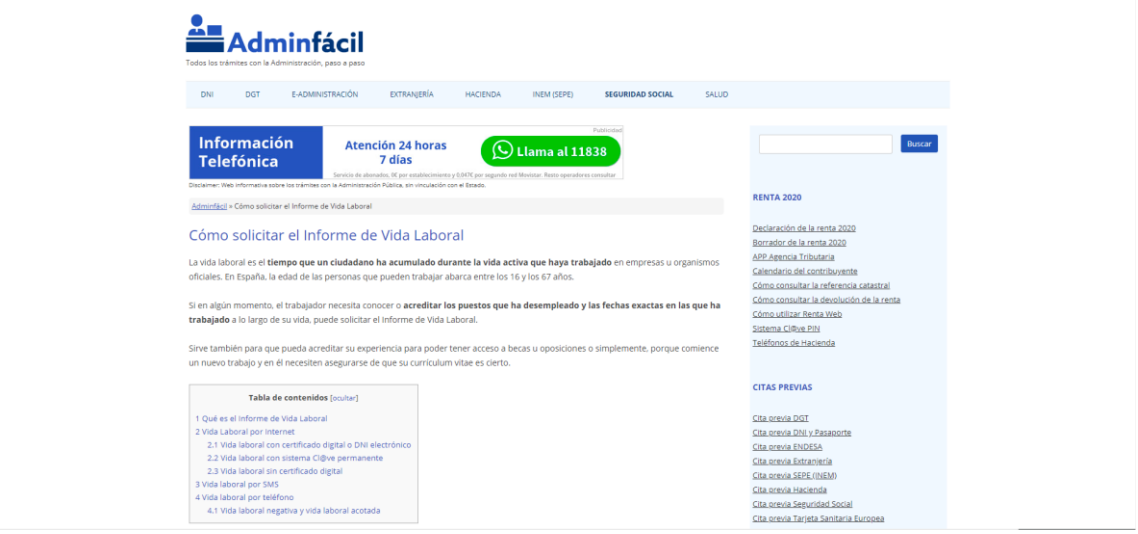


Ilustración 13 Adminfacil.es vida laboral [21]

Ilustración 14seg-social vida laboral [22]

- Burbuja.info

Ilustración 15Burbuja.info incorrecciones declaración [23]

Ilustración 16agenciatributaria.com incorrecciones declaración [18]

El spider analizará dichas URLs y almacenará la página web en local. En el análisis lo que se busca es obtener el título de la página web, y el texto de la misma. A su vez, el spider busca URL a páginas web dentro de dicha página web y los analiza a su vez. Este proceso será capado para que los dominios a los que pueda acceder para analizar la página

web pertenezcan a las URLs inicialmente aportadas, debido a que, si no se limita el alcance del spider, este comienza un proceso iterativo en el cual accede a multitud de páginas web las cuales no ofrecen ningún tipo de información útil para el programa.

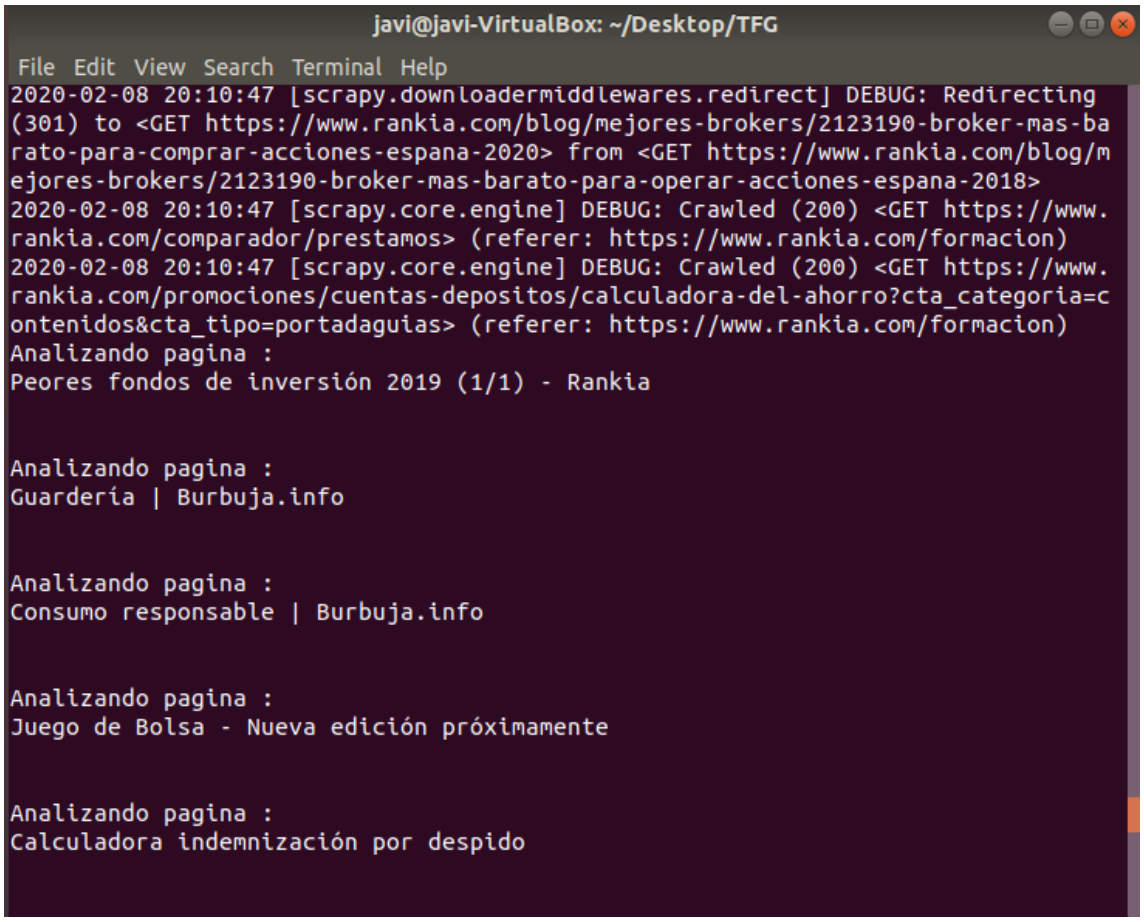


Ilustración 17Obtención páginas spider

En la siguiente tabla se muestran el número de páginas recolectadas por el spider para cada uno de los portales que han sido incluidos en los dominios permitidos por el spyder.

PORTAL	NÚMERO DE PÁGINAS WEB
RANKIA.COM	1808
BURBUJA.INFO	226
ADMINFACIL.COM	999
COMOSETRAMITA.COM	1250
TOTAL	**4283**

Tabla 2Páginas analizadas spider

Este proceso conllevará varios minutos en la realización de la recuperación de las distintas páginas, por lo que se ha decidido realizarlo como un proceso previo independiente al uso de la herramienta, independiente, el cual se ejecuta mediante un *cronjob* todos los días a las 00:00.

El motivo de esta decisión ha sido que el número de páginas creadas en cada uno de los portales a diario no es muy elevado, y en un entorno real en el que personas de España utilizasen este servicio, sería una hora que no afectaría al servicio y por lo tanto permitiría mantener actualizada la herramienta.

Para el cálculo de la similitud de la página web a analizar con las páginas en los portales, en primer lugar, se tokeniza el título de la página web a analizar ya que se utilizará como consulta y los datos de cada una de las páginas web ofrecidas por los portales. A continuación, se procede a realizar el proceso del cálculo de la similitud mediante TFIDF.

Una vez calculado TFIDF frente a todas las páginas, se obtiene el mayor valor de todos ellos, es decir la página web con una similitud mayor que todas las demás y se almacena en la base de datos.

A continuación, se puntúa el número de imágenes encontrados en la página a analizar, ya que estas facilitan la comprensión de textos dando en muchas ocasiones una ayuda visual, sin embargo, se ha determinado que, si se el porcentaje de imágenes supera el 50% del tamaño de la página, esto hace que dichas imágenes saturen la página y no permitan la correcta comprensión del texto.

A continuación, se procede a analizar el texto. Se analizan cada una de las palabras con respecto a si dichas palabras son tecnicismos, es decir, pertenecen a un campo concreto y su familiaridad.

Para el cálculo de la familiaridad, se parte de la base de que los portales que ofrecen información transcrita de las páginas administrativas web del Estado utilizan un vocabulario con una mayor familiaridad para el usuario.

Partiendo de esta premisa, lo que se ha calculado es la frecuencia media de los términos en los documentos, es decir, el IDF medio, utilizando como corpus el conjunto de textos obtenidos de los distintos portales que ofrecen transcripciones de las páginas administrativas web del Estado.

A su vez, para el cálculo de los tecnicismos, se ha comprobado que las palabras consideradas como tecnicismos tienen un uso menor en los textos de los portales que ofrecen transcripciones de las páginas administrativas web del Estado, por lo tanto, se ha establecido que los términos con un IDF menor que un umbral establecido a partir de los valores IDF de varios tecnicismos obtenidos de varios textos serán entendidos como tecnicismos.

En este caso se incluyen a su vez como tecnicismos toda aquella palabra que sea poco frecuente en estas páginas, por lo que la acepción de tecnicismo no es la más acertada.

Por último, se cuenta el número de palabras de cada una de las frases y el número de palabras en cada uno de los párrafos dando un valor numérico calculado a partir del número de palabras total del texto y palabras de cada uno de los párrafos.
Una vez calculados todos estos parámetros, se procede a la generación de la valoración de la página web mediante la inteligencia artificial elegida.

En este caso, se ha decidido realizar la evaluación de la página web mediante una red neuronal [13]. La decisión de utilizar una red neuronal es debido a su capacidad como aproximador universal, es decir, es capaz de obtener valoraciones para todo tipo de inputs que se le introduzca, por lo que, al poder cambiarse o introducirse nuevos parámetros de análisis de la comprensibilidad de la página web, este tipo de inteligencia artificial nos permitirá añadirlos.

Un ejemplo de una aproximación mediante un perceptron multicapa se encuentra en la Revista Internacional de Métodos Numéricos para Cálculo y Diseño en Ingeniería [17]

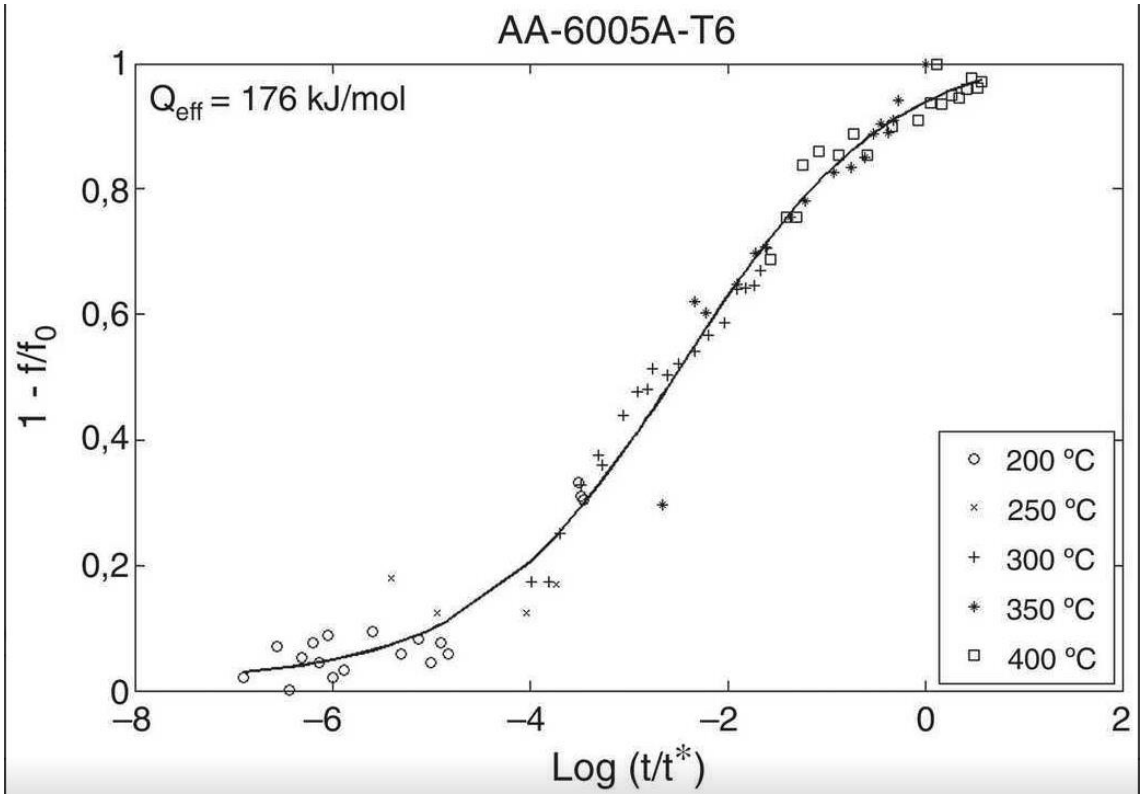


Ilustración 18Ejemplo perceptron [18]

En la imagen podemos observar como dadas una serie de muestras (los distintos círculos, cruces y cuadrados), el perceptron (la línea continua) se ajusta a los datos, pudiendo así aproximar la salida de cualquier entrada que se le introduzca.

La eficiencia del *perceptron* queda demostrada en el artículo escrito por Adil Tannouche, , Khalid Sbai , Youssef Ounejjar y Abdelali Rahmani, en el cual se utiliza un perceptron con el fin de realizar la gestión de semillas de cebolla mediante esta red neuronal. [13]

A su vez, la facilidad de implementación de la inteligencia artificial es un punto a favor a la hora de la seleccionarla, debido a que, al desarrollarse sobre Python, existen múltiples librerías que implementan la red neuronal perceptron.

Por último, destacar la eficiencia de dicha inteligencia artificial, ya que el tiempo de procesamiento de la red neuronal a la hora de obtener valoraciones de páginas web es muy corto comparado con otras inteligencias artificiales.

Esta valoración generada por la inteligencia artificial será almacenada en la base de datos junto con toda la información sobre la página web en la colección que contiene las páginas valoradas.

En este punto, en la interfaz web se mostrará la valoración final de la página web obtenida de la base de datos.

En el caso de que la página web ya haya sido analizada con anterioridad por un usuario con la misma edad que el actual, el proceso será mucho más breve, ya que al haber almacenado cada una de las páginas web y su valoración, únicamente se comprobará si la página web ha sido almacenada en la base de datos con anterioridad con un análisis sobre un usuario con la misma edad y se le mostrará esa valoración en la interfaz web.

DIAGRAMA DE DESPLIEGUE

En este apartado se definirá la disposición de los distintos componentes del sistema y como estarán conectados cada uno de ellos.

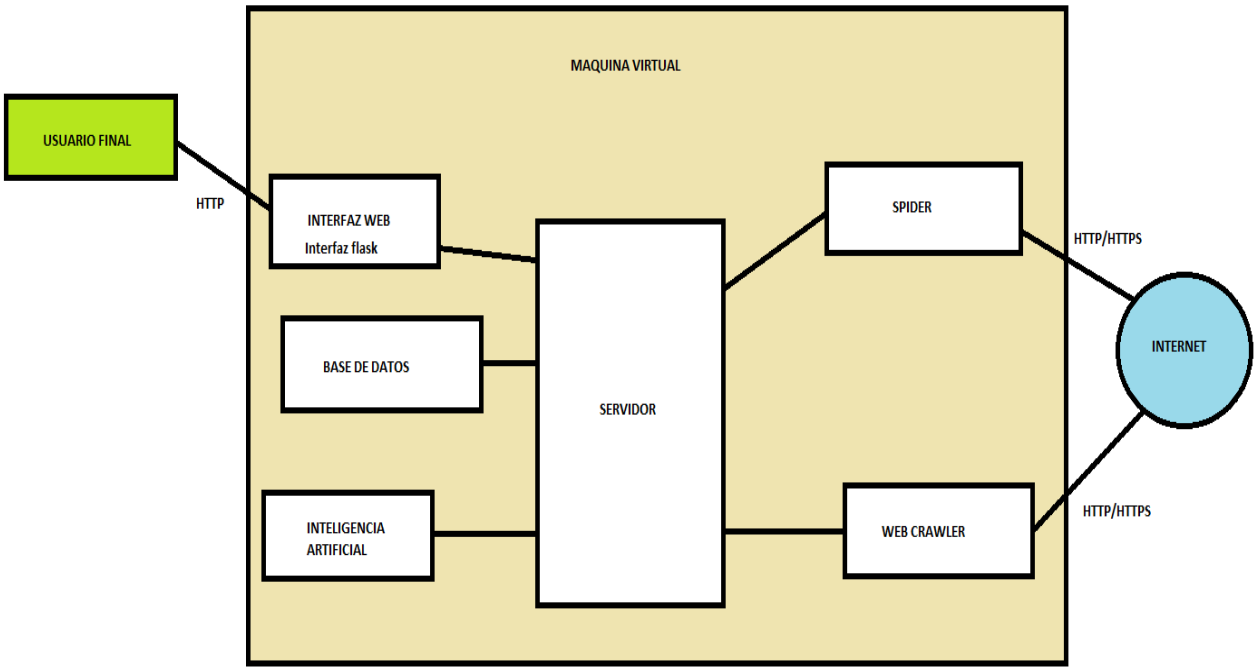


Ilustración 19Diagrama de despliegue

En la imagen podemos observar, en primer lugar que todo el sistema se encontrará situado en una máquina virtual, la cual en un primer lugar es Oracle VirtualBox [12], pero sin embargo gracias a la portabilidad de esta herramienta, es posible la migración del sistema a otros entornos.

Los distintos componentes dentro de la máquina virtual son llamados y ejecutados por el servidor en función de las llamadas al sistema que reciba mediante la interfaz web.

La interfaz web es el enlace entre el sistema y el usuario, el cual permitirá al usuario realizar peticiones sobre el análisis de las distintas páginas a la vez que permitirá al sistema mostrar los resultados de las evaluaciones de las páginas.

A su vez, el spider, el cual se encargará de recolectar las páginas de los distintos portales que ofrecen transcripciones de las páginas del estado simplificadas como el web crawler, encargado de la obtención de los distintos parámetros de la página web a analizar necesitan de acceso a la red o internet. Por ello en el diagrama de despliegue se ha visto necesario remarcar dicha conexión. Esta conexión se hará mayoritariamente mediante *https*, sin embargo, se han encontrado una serie de páginas que aún están en *http* por lo que se ha decidido soportar ambos estándares de conexión.

ENTORNO Y HERRAMIENTAS

En este apartado se define el entorno en el cual se ha desarrollado el sistema y las distintas herramientas utilizadas para el desarrollo del mismo.

ENTORNO DE DESARROLLO

Con respecto al entorno, toda la herramienta ha sido creada sobre una máquina virtual montada en un Windows 10 con el programa Oracle VM VirtualBox [14].

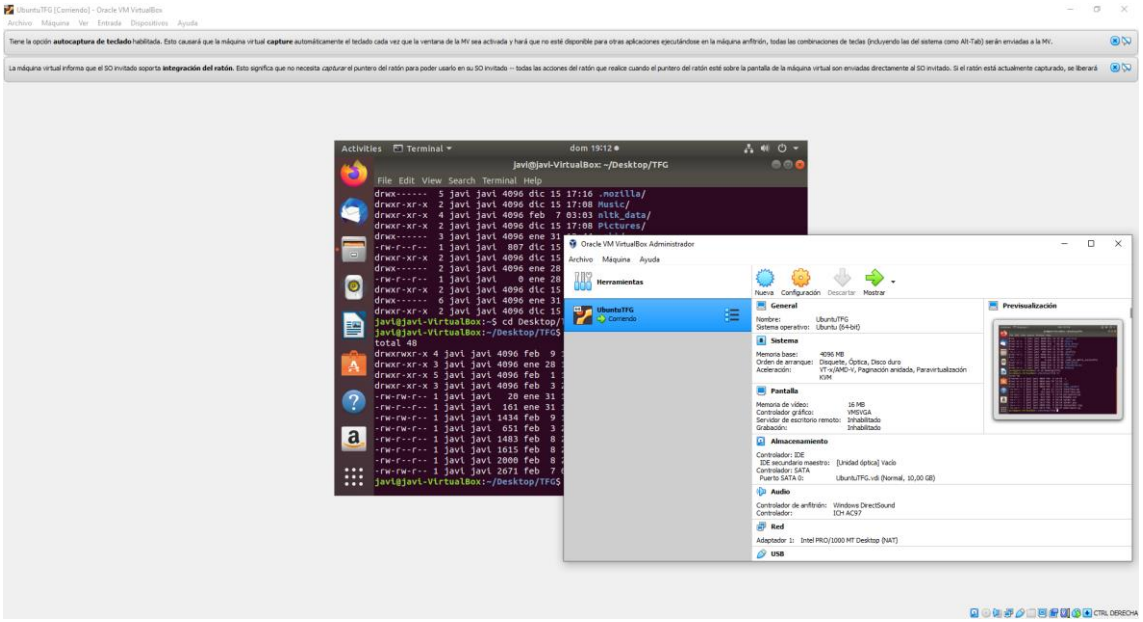


Ilustración 20Entorno VirtualBox

VirtualBox es una potente herramienta de virtualización la cual te permite extender las capacidades de un equipo haciéndolo soportar múltiples sistemas operativos con una única instalación en el equipo.

A su vez, se ha de destacar el poder controlar los recursos utilizados por las distintas máquinas virtuales en el sistema anfitrión, a la vez que permite la integración de dichas máquinas virtuales con el sistema anfitrión u otros sistemas.

Se ha de destacar a su vez la portabilidad de los distintos sistemas creados en este entorno a otros equipos o incluso a otras herramientas, permitiendo así el desarrollo en diversos entornos.

En esta máquina virtual, se ha instalado la distribución del sistema operativo Linux llamada Ubuntu de 64-bits [15].

Se ha decidido utilizar Linux en un primer lugar por contener un compilador de Python por defecto y por la familiaridad con el mismo a la hora del desarrollo de herramientas en lenguajes como C++ o Python [16].

Al mismo tiempo otra de las razones por las cuales se ha elegido este sistema operativo es la facilidad de automatización de pruebas. Esto se debe a la programación en Bash, ya que permite multitud de acciones y repetición de las mismas mediante simples comandos.

La elección de la distribución de Linux Ubuntu se debe a la facilidad de encontrar dicha distribución y a la familiaridad con la misma. Se barajó la utilización de distintas distribuciones, debido a la alta capacidad que requiere Ubuntu, ya que otras distribuciones como BunsenLabs [17] no tienen una interfaz gráfica tan elaborada como la de Ubuntu y permite distribuir los recursos no utilizados en dicha interfaz en recursos de procesamiento.

LENGUAJE DE PROGRAMACIÓN E INTERFACES DE DESARROLLO

Como lenguaje de programación a la vez que compilador se ha utilizado Python3 debido a que los pequeños cambios entre la versión Python2.7 y la versión Python3 permite saber programar en ambas versiones solo aprendiendo una. A su vez, con estos cambios añadidos en Python3 hacen que sea mucho más simple y ágil la programación que su predecesora, en este caso Python2.7. Por otra parte, Python ha conseguido ser uno de los lenguajes con un mayor prestigio en el ámbito del análisis lingüístico, debido al gran número de librerías desarrolladas para este lenguaje, en las que hay que destacar la librería NLTK tools [18], la cual es una de las colecciones de procesamiento del lenguaje más antiguas y por lo tanto más completas al igual que está ampliamente instaurada en la docencia de dicha materia.

Principalmente la herramienta ha sido desarrollada mediante el editor de texto de Linux, ya que, al utilizar la distribución de Linux Ubuntu, el editor de texto contiene una interfaz gráfica que permite un desarrollo más ágil.

A su vez, con el fin de agilizar la creación de la herramienta y la facilidad de poder realizar desarrollos o pruebas en distintos entornos, se ha utilizado la interfaz web de Jupiter notebook [19].

Jupiter es una herramienta que permite desarrollar y testear código en diversos lenguajes, entre los que se encuentra Python, tanto a como de forma local, como online. Esto ha permitido el desarrollo de la herramienta con una mayor facilidad, ya que ha permitido tener el código en múltiples equipos. A su vez, no es necesario tener que instalar ningún software adicional independientemente del sistema operativo en el que se ejecute esta herramienta, ya que Jupiter se ejecuta sobre el buscador.

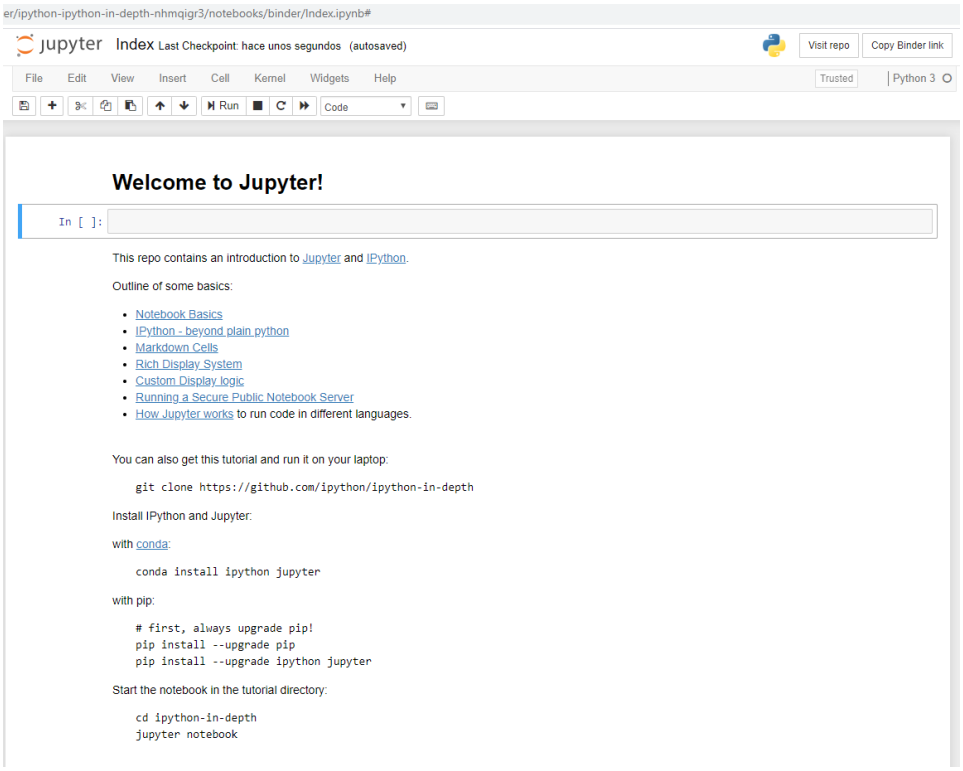


Ilustración 21Entorno Jupiter Notebook

Adicionalmente, se han utilizada numerosas librerías para realizar las funcionalidades requeridas. En las próximas secciones se facilitan detalles sobre estas librerías.

LIBRERÍAS UTILIZADAS

TRATAMIENTO DE PAGINAS WEB

En primer lugar, para poder trabajar con URLs y descargarlas para su posterior análisis se ha utilizado la librería urllib [20].

El módulo de esta librería llamado urllib.request permite hacer peticiones a distintos servicios web o páginas web con el fin de obtener la información de la misma como una variable que se puede tratar.

Otra de las librerías utilizadas en el desarrollo de la herramienta ha sido la librería BeautifulSoup [21]. Esta librería permite la obtención de datos a partir de un HTML para su procesamiento, por lo que, junto con la librería anteriormente mencionada, nos permite obtener toda la información de la página web.

CALCULO DE PARÁMETROS

Para el análisis de la similitud entre documentos se ha utilizado la librería de NLTK llamada wordnet [28]. Esta librería permite la creación de un conjunto de palabras para cada uno de los textos a analizar llamados *synsets* y a su vez calcular la similitud de ambos textos.

A su vez, para el cálculo de las medidas TF, IDF y TFIDF se ha utilizado a su vez otra librearía de NLTK llamada *TFIDFvectorizer*. Esta librería permite el cálculo de dichos parámetros dado un conjunto de documentos.

ALMACENAMIENTO DE DATOS

Como base de datos se ha decidido utilizar MongoDB [22]. A diferencia de las bases de datos comúnmente vistas, MongoDB es una base de datos NoSQL documental. Las bases de datos documentales son aquellas las cuales almacenan los datos en documentos. Estos

documentos son conjuntos de datos semiestructurados los cuales permiten una mayor flexibilidad a la hora de su almacenamiento que las bases de datos relacionales.

Al ser necesario almacenar una cantidad de datos aleatoria, ya que por ejemplo el número de palabras por página web es variable, una base de datos no nos permitiría almacenar los datos de una forma que fuera cómoda para su posterior análisis.

CALCULO DE VALORACIÓN MEDIANTE RED NEURONAL

Con respecto a la inteligencia artificial, se ha decidido utilizar la red neuronal llamada perceptron. Al trabajar en Python, este lenguaje nos ofrece una librería llamada sklearn [23], la cual contiene una librería específica para el perceptron, la cual nos permite entrenar, ajustar y obtener valoraciones mediante esta red neuronal.

Para el tratamiento de los datos a la hora de introducirse a la red neuronal, en este caso al perceptron, se ha utilizado la librería de Python llamada Numpy [24]. Esta librería permite a partir de una serie de variables, crear una matriz donde almacenar los datos que se van a introducir como parámetros.

En un primer lugar, se utilizó la librería JSON de Python para trabajar sobre los documentos devueltos por MongoDB, sin embargo, estos documentos son variables de tipo dictionary, las cuales se pueden tratar de manera sencilla sin necesidad de ninguna librería adicional, por lo que se descartó la utilización de JSON.

INTERFAZ GRÁFICA

Por último, para el desarrollo de la interfaz web se ha utilizado JavaScript [25]. Un lenguaje de programación sencillo y ágil el cual permite utilizar distintas librerías para realizar llamadas a programas escritos en Python.

A su vez, la librería Flask de Python [26], permite la creación de servicios web de manera simple donde poder crear una interfaz web a la vez que realizar llamadas a métodos creados en Python y poder mostrar los resultados de los distintos métodos en la interfaz web.

Para el debugging de la interfaz gráfica, se ha utilizado las herramientas proporcionadas por Google para el desarrollo web con el fin de localizar errores en el código o elementos que se mostraban incorrectamente.

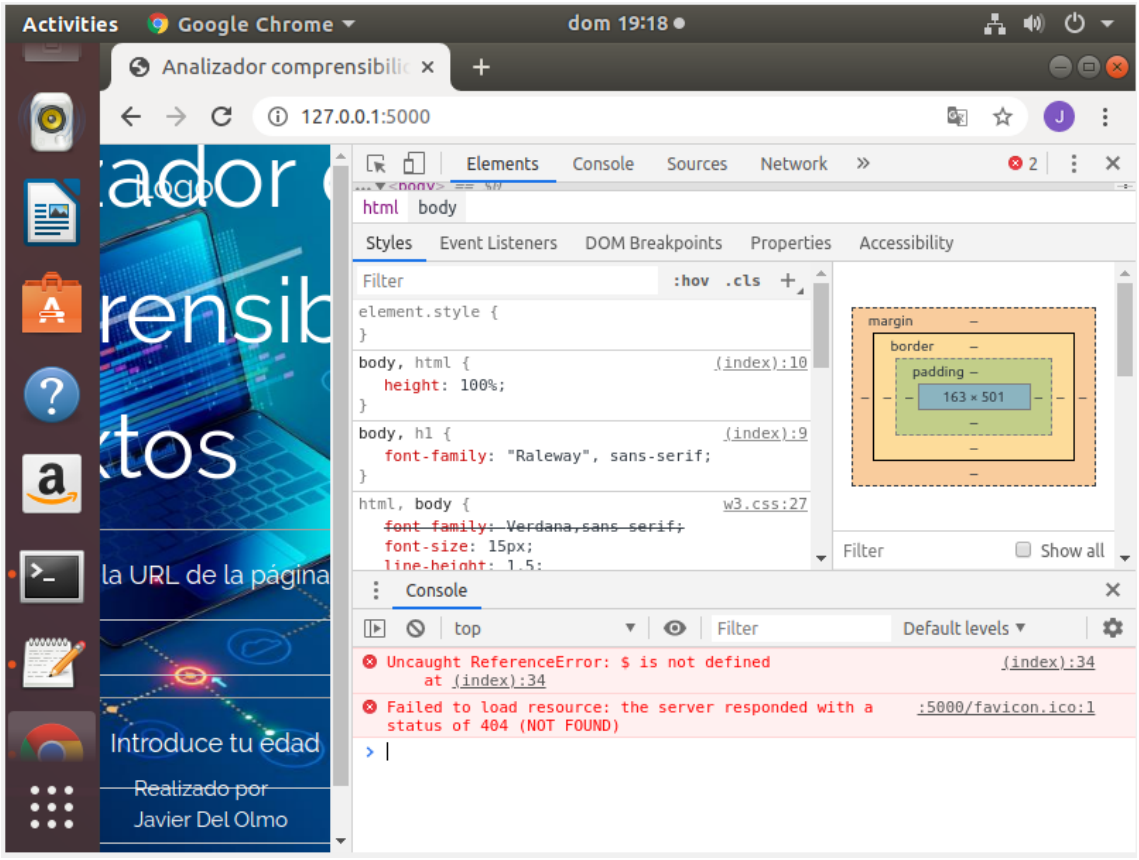


Ilustración 22Developer tools chrome

Esta herramienta ofrece una interfaz la cual permite localizar cada elemento de la página web fácilmente y mostrando los errores de la página en una consola como se puede ver en la imagen.

OBTENCIÓN DE DOCUMENTOS DE ENTRENAMIENTO

Para el entrenamiento de la inteligencia artificial ha sido necesario obtener un conjunto de páginas web de procedimientos web del Estado analizadas por distintos usuarios.
Para ello, en primer lugar, se ha preseleccionado un conjunto de 20 páginas de procedimientos administrativos web del Estado. Estas 20 páginas elegidas han sido enviadas a 20 personas distintas, de distintas edades y distinta formación académica.

A estas 20 personas se les ha enviado un documento en formato Excel, en el cual se indicaba una lista de 20 páginas web y estas personas devolvían el documento indicando su edad, y una valoración entre 1 y 100 de la comprensibilidad de cada una de las distintas páginas web listadas.

Estas personas han analizado la comprensibilidad de cada una de las páginas dando una valoración de la misma y devolviendo a su vez su edad.

Estos análisis han sido almacenados en la base de datos y utilizados como conjunto de entrenamiento de la herramienta.

Cabe destacar que, en los resultados obtenidos, para gente con una edad menor de 30 años la valoración obtenida en los textos no variaba mucho para personas de una misma edad, sin embargo, en edades superiores a 50 años, la formación académica de las personas que realizaron estas valoraciones es bastante distante por lo que las valoraciones varían bastante entre unas personas y otras.

ANÁLISIS

En este apartado se definirán los casos de uso de la herramienta desarrollada, el diagrama de clases de la base de datos y los requisitos tanto funcionales como no funcionales asociados a este proyecto.

ESTADO INICIAL DEL PROYECTO

Actualmente el campo de la comprensibilidad de los textos es un campo estudiado para cada uno de los distintos lenguajes por diversos proyectos desarrollados, tanto por particulares, como por distintas instituciones como pueden ser las universidades [5].

Sin embargo, existen escasas herramientas de análisis en español de páginas web completas. A su vez, la mayoría de las herramientas desarrolladas con el fin de analizar un texto utilizan funciones o métodos lineales, por lo que existen escasos estudios centrados en el castellano, y más aún en esta tipología documental.

CASOS DE USO

CU-01	Analizar página web del estado	
Dependencias	RF-01: Valorar página web RF-02: El usuario obtiene mediante la interfaz web el valor de la comprensibilidad de la página RF-06: El usuario introducirá los parámetros desde la página web RNF-01: La página debe ser sencilla y fácil de manejar RNF-02: El sistema estará disponible en cualquier momento para realizar valoraciones RNF-03: El tiempo de respuesta del sistema no será superior a 5 minutos.	
Precondición	El usuario deberá acceder a la herramienta y tener la URL de la página web a analizar, no perteneciendo esta al estado	
Descripción	El usuario introducirá la URL en el cuadro de texto dispuesto para ello y la página devolverá a continuación el valor de la comprensibilidad de dicho texto.	
Secuencia normal	Paso	Acción
	1	El usuario accede a la herramienta mediante la URL de la misma
	2	El usuario introduce en el cuadro de texto la página web a analizar
	3	El sistema realiza los cálculos pertinentes para el cálculo de la comprensibilidad de la página.
	4	El sistema muestra en la interfaz web el resultado del valor de la comprensibilidad de la página web.

Postcondición	El usuario lee de la página web el valor de la comprensibilidad de la página web a analizar		
Excepciones	Paso	Acción	
1	1	El usuario inserta una página web que no pertenece a una web del estado	
		E.1	El sistema muestra un mensaje advirtiendo que la página web no pertenece al estado y por lo que no es posible realizar dicho análisis
		E.2	El sistema muestra por la interfaz un botón que permite analizar otra página.
		E.3	Tras pulsar el botón, el usuario vuelve a la página inicial
Comentarios	La página solo permite el análisis de páginas web pertenecientes al estado que definan procesos burocráticos. Las páginas que no pertenecen a este grupo no serán analizadas con el fin de evitar desajustes en el entrenamiento previo de la inteligencia artificial.		
2	1	El usuario inserta una edad no válida	
		E.1	El sistema muestra un mensaje advirtiendo que la edad introducida por el usuario no es válida y por lo que no es posible realizar dicho análisis

		E.2	El sistema muestra por la interfaz un botón que permite analizar otra página.
		E.3	Tras pulsar el botón, el usuario vuelve a la página inicial y la edad del usuario
Comentarios	El sistema no permite el análisis de páginas web en las que el usuario haya introducido un valor para la edad superior a 120 o inferior a 0. Las páginas que el valor de la edad no cumpla estas condiciones no serán analizadas con el fin de evitar desajustes en el entrenamiento previo de la inteligencia artificial.		

Tabla 3Caso de uso 1

CU-02	Mostrar los parámetros de análisis utilizados para la valoración de una página web.
Dependencias	RF-01: Valorar página web RF-03: El sistema muestra por la interfaz web los parámetros utilizados en la valoración de la página web RF-06: El usuario introducirá los parámetros desde la página web RNF-01: La página debe ser sencilla y fácil de manejar RNF-02: El sistema estará disponible en cualquier momento para realizar valoraciones RNF-03: El tiempo de respuesta del sistema no será superior a 5 minutos.
Precondición	El usuario deberá acceder a la herramienta y tener la URL de la página web a analizar, perteneciendo esta al estado

Descripción	El usuario introducirá la URL en el cuadro de texto dispuesto para ello y pulsará el botón de obtener parámetros. La página devolverá a continuación los parámetros utilizados para la valoración de la página.		
Secuencia normal	Paso	Acción	
	1	El usuario accede a la herramienta mediante la URL de la misma	
	2	El usuario introduce en el cuadro de texto la página web a analizar y su edad.	
	3	El sistema realiza los cálculos pertinentes para el cálculo de la comprensibilidad de la página.	
	4	El sistema muestra en la interfaz web los parámetros obtenidos para el cálculo de la comprensibilidad de la página web.	
Postcondición	El usuario obtiene mediante la interfaz web los parámetros utilizados para el análisis de la página web.		
Excepciones	Paso	Acción	
	1	El usuario inserta una página web que no pertenece a una web del estado	
		E.1	El sistema muestra un mensaje advirtiendo que la página web no pertenece al estado y por lo que no es posible realizar dicho análisis
		E.2	El sistema muestra por la interfaz un botón que permite analizar otra página.

		E.3	Tras pulsar el botón, el usuario vuelve a la página inicial
Comentarios	La página solo permite el análisis de páginas web pertenecientes al estado que definan procesos burocráticos. Las páginas que no pertenecen a este grupo no serán analizadas con el fin de evitar desajustes en el entrenamiento previo de la inteligencia artificial.		
2	1	El usuario inserta una edad no válida	
		E.1	El sistema muestra un mensaje advirtiendo que la edad introducida por el usuario no es válida y por lo que no es posible realizar dicho análisis
		E.2	El sistema muestra por la interfaz un botón que permite analizar otra página.
		E.3	Tras pulsar el botón, el usuario vuelve a la página inicial y la edad del usuario
Comentarios	El sistema no permite el análisis de páginas web en las que el usuario haya introducido un valor para la edad superior a 120 o inferior a 0. Las páginas que el valor de la edad no cumpla estas condiciones no serán analizadas con el fin de evitar desajustes en el entrenamiento previo de la inteligencia artificial.		

Tabla 4Caso de uso 2

DIAGRAMA DE CLASES

Con respecto al diagrama de clases, al haberse utilizado una base de datos en MongoDB, no se ha visto conveniente la realización de un diagrama de clases convencional, por lo que se ha decantado por el uso de UML con el fin de mostrar la estructura de los documentos y los distintos elementos que contiene cada uno de los datos recolectados de la página web.

Colección	Paginas_a_valorar
Columna	Descripción
_ID	Identificador unico
url	URL de la página
Titulo	Título de la página
edad	Edad del usuario
palab_div	Media de palabras por div
palab_parraf	Media de palabras por parrafo
Fam_text	Familiaridad media del texto
Sim_text	Similitud maxima con los portales
NumTecn	Número de tecnicismos
TFIDF	Calculo TFIDF de la página
IDF_Total	IDF total de los terminos

Tabla 5Diagrama de clases página valorada

Colección	Paginas_valoradas
Columna	Descripción
_ID	Identificador unico
url	URL de la página
Titulo	Título de la página
edad	Edad del usuario
palab_div	Media de palabras por div
palab_parraf	Media de palabras por parrafo
Fam_text	Familiaridad media del texto
Sim_text	Similitud maxima con los portales
NumTecn	Número de tecnicismos
TFIDF	Calculo TFIDF de la página
IDF_Total	IDF total de los terminos
Values	Valoración de la comprensibilidad de la página

Tabla 6Diagrama de clases página valorada

REQUISITOS FUNCIONALES

A continuación, se definen los requisitos funcionales definidos para dicha aplicación.

RF-01	**Analizar página**
Prioridad	Alta
Dependencias	RF-02, RF-03, RF-04, RNF-01, RF-06
Descripción	Dada una página perteneciente al estado, el sistema será capaz de obtener los parámetros necesarios para el cálculo de la comprensibilidad del texto.
Pruebas	PF-01, PF-02

Tabla 7 Requisito funcional 1

RF-02	**Mostrar resultados de la página analizada**
Prioridad	Alta
Dependencias	RF-02, RF-03, RF-04, RNF-01, RF-06
Descripción	Tras el análisis de una página web proporcionada por un usuario, el sistema será capaz de mostrar por pantalla la valoración de la página web.
Pruebas	PF-02

Tabla 8 Requisito funcional 2

RF-03	**Mostrar parámetros de la página**
Prioridad	Alta
Dependencias	RF-02, RF-03, RF-04, RNF-01, RF-06
Descripción	El sistema podrá mostrar por pantalla al usuario los parámetros utilizados para el cálculo de la comprensibilidad de la página.
Pruebas	PF-05

Tabla 9Requisito funcional 3

RF-04	**Error en caso de página web no perteneciente a procesos administrativos web**
Prioridad	Alta
Dependencias	RF-02, RF-03, RF-04, RNF-01, RF-06
Descripción	El sistema devolverá un error en caso de que la URL introducida por el usuario no pertenezca a un procedimiento administrativo web del estado.
Pruebas	PF-03

Tabla 10Requisito funcional 4

RF-05	**Error en caso de valor de edad no válida**
Prioridad	Alta
Dependencias	RF-02, RF-03, RF-04, RNF-01, RF-06
Descripción	El sistema devolverá un error en caso de que el valor de la edad introducida por el usuario no se encuentre entre el 0 y el 120
Pruebas	PF-04

Tabla 11Requisito funcional 4

RF-06	**El sistema permitirá introducir parámetros mediante la interfaz web**
Prioridad	Alta
Dependencias	RF-02, RF-03, RF-04, RNF-01, RF-06
Descripción	El sistema proporcionará al usuario una interfaz web mediante la cual el usuario podrá insertar la página a analizar y la edad del usuario
Pruebas	PF-01, PF-02, PF-03, PF-04, PF-05

Tabla 12Requisito funcional 4

REQUISITOS NO FUNCIONALES

En las siguientes tablas se definen los requisitos no funcionales que afectan al sistema.

RNF-01	**Disponibilidad del sistema**
Prioridad	Alta
Dependencias	RF-01
Descripción	El sistema estará disponible para la evaluación de páginas web las 24 horas del día
Pruebas	PNF-01

Tabla 13Requisito no funcional 2

RNF-02	**Tiempo de respuesta del sistema**
Prioridad	Alta
Dependencias	RF-01
Descripción	El tiempo de procesado y su posterior generación de la valoración de la página web no será superior a 5 minutos.
Pruebas	PNF-02

Tabla 14Requisito no funcional 4

RNF-03	**Datos de portales actualizados**
Prioridad	Alta
Dependencias	RF-01
Descripción	Los datos de las páginas web de los portales que proporcionan información para el cálculo de la comprensibilidad de la página serán actualizadas una vez al día
Pruebas	PNF-03

Tabla 15Requisito no funcional 5

PRUEBAS

En este apartado se define cada una de las distintas pruebas realizadas para la comprobación de los distintos requisitos definidos anteriormente.

Se va a dividir las pruebas en pruebas de requisitos funcionales y pruebas de requisitos no funcionales.

PRUEBAS REQUISITOS FUNCIONALES

PF-01	**Analizar página**
Prioridad	Alta
Dependencias	RF-01
Descripción	Dada una URL proporcionada por el usuario, la cual no se encuentre en el corpus de entrenamiento de la inteligencia artificial y su edad el sistema evaluará la página proporcionada.
Especificaciones de entrada	URL perteneciente a un dominio de procesos administrativos web del estado y la edad del usuario que ha solicitado la petición.
Especificaciones de salida	Valoración almacenada en el sistema
Resultado	Éxito

Tabla 16Prueba funcional 1

PF-02	**Mostrar página analizada**
Prioridad	Alta
Dependencias	RF-01, RF-02
Descripción	Dada una URL proporcionada por el usuario y una edad el sistema evaluará la página proporcionada y mostrará al usuario la valoración obtenida para dicha página mediante la interfaz web.
Especificaciones de entrada	URL perteneciente a un dominio de procesos administrativos web del estado y la edad del usuario que ha solicitado la petición.
Especificaciones de salida	Valoración de la página web mediante la interfaz web.
Resultado	Éxito
Adjuntos adicionales a las pruebas	

Ilustración 23Prueba funcional 2

Tabla 17Prueba funcional 2

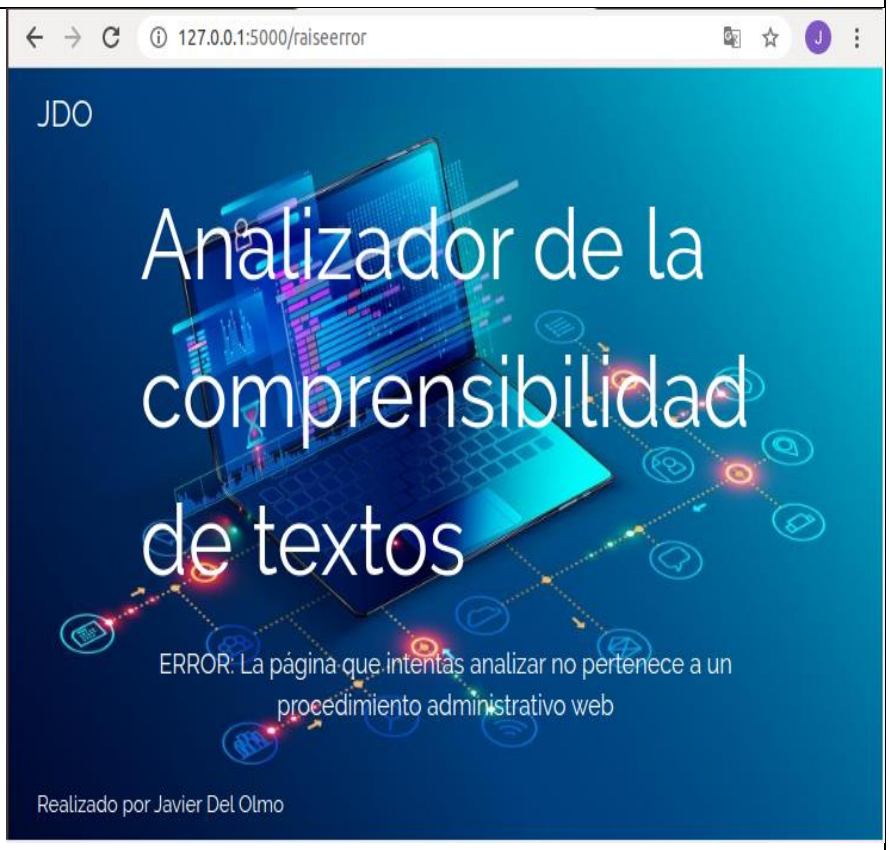
PF-03	**Error al introducir una página del estado no perteneciente a procesos administrativos web del estado**
Prioridad	Alta
Dependencias	RF-04
Descripción	Si la URL de la página web a valorar no se encuentra en uno de los dominios definidos como páginas web de procesos administrativos del estado, la herramienta mostrará en la interfaz web un error el cual indicará el motivo del error al usuario.
Especificaciones de entrada	URL no perteneciente a un dominio de procesos administrativos web del estado
Especificaciones de salida	Mensaje de error mostrado mediante la interfaz web indicando que no dicha página no pertenece a un dominio de procesos administrativos web del estado.
Resultado	Éxito
Adjuntos adicionales a las pruebas	

Ilustración 24prueba funcional 3

Tabla 18Prueba funcional 3

PF-04	**Error al introducir una edad no valida.**
Prioridad	Alta
Dependencias	RF-05
Descripción	Si la edad introducida por el usuario no es válida, el sistema mostrará por la interfaz web un mensaje de error indicando que el valor de la edad no es válido
Especificaciones de entrada	Valor de la edad inferior a 0 o superior a 120
Especificaciones de salida	Mensaje de error indicando que la edad introducida no es válida.
Resultado	Éxito
Adjuntos adicionales a las pruebas	 *Ilustración 25prueba funcional 4*

Tabla 19Prueba funcional 4

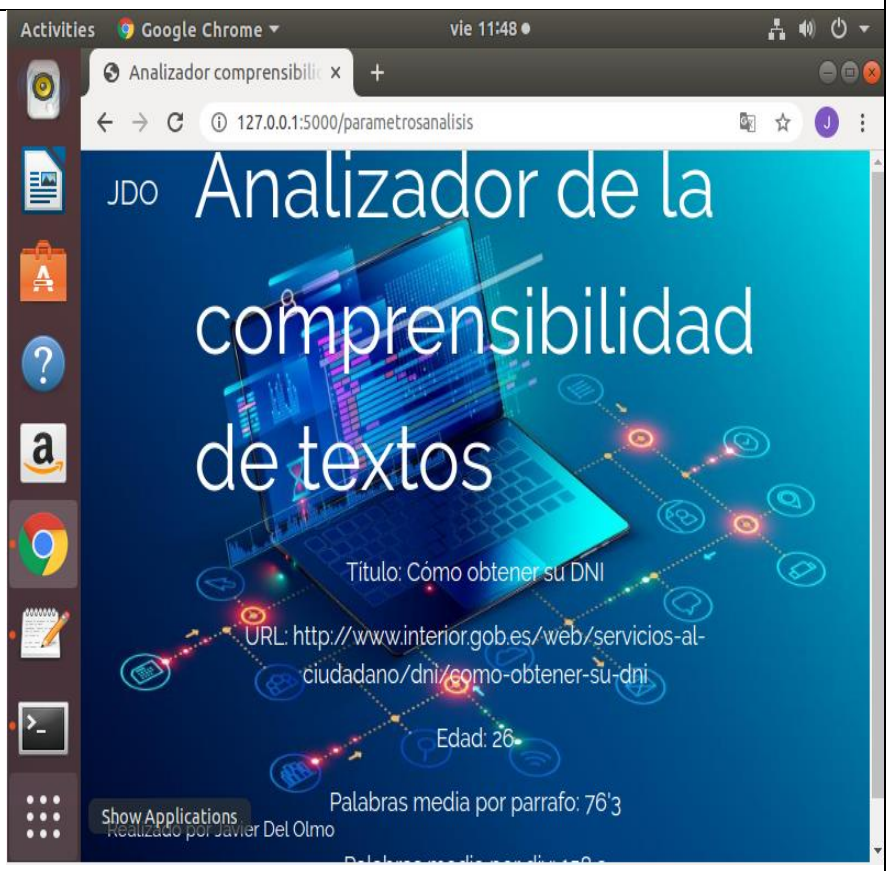
PF-05	**Mostrar parámetros de análisis.**
Prioridad	Alta
Dependencias	RF-02
Descripción	El usuario solicita al sistema que muestre los parámetros utilizados a la hora de analizar una página web indicada por el usuario
Especificaciones de entrada	URL de la página web a analizar y la edad del usuario
Especificaciones de salida	Parámetros utilizados en el análisis.
Resultado	Éxito
Adjuntos adicionales a las pruebas	

Ilustración 26prueba funcional 5

Tabla 20Prueba funcional 5

PRUEBAS REQUISITOS NO FUNCIONALES

PNF-01	**Comprobar disponibilidad del sistema**
Prioridad	Alta
Dependencias	RNF-01
Descripción	El sistema se mantendrá encendido durante 3 días seguidos y comprobando que sigue analizando las páginas solicitadas cada hora.
Especificaciones de entrada	Solicitud de valoración de un procedimiento web del estado mediante un script.
Especificaciones de salida	Respuesta del sistema.
Resultado	Éxito

PNF-02	**Comprobar tiempo de respuesta de la página**
Prioridad	Media
Dependencias	RNF-02
Descripción	Dada una URL proporcionada por el usuario y una edad el sistema evaluará la página proporcionada en menos de 5 minutos.
Especificaciones de entrada	URL perteneciente a un dominio de procesos administrativos web del estado y la edad del usuario que ha solicitado la petición.
Especificaciones de salida	Tiempo en procesar la valoración y mostrarla al usuario.
Resultado	Éxito

PNF-03	**Comprobar actualización diaria de los portales**
Prioridad	Media
Dependencias	RNF-03
Descripción	El sistema cada día a las 00:00 realizará un rastreo de nuevas páginas web de los portales que transcriben páginas web del estado buscando nuevas entradas.
Especificaciones de entrada	Ninguna
Especificaciones de salida	Actualización paginas aportadas por los portales
Resultado	Éxito

DESARROLLO

El desarrollo se ha realizado implementando cada uno de los distintos componentes del sistema de manera individual y por último ajustando cada uno de ellos y uniéndolos, conformándose así el sistema final.

LENGUAJE DE PROGRAMACIÓN

El sistema se ha desarrollado sobre Python debido a la multitud de librerías que han sido desarrolladas para dicho lenguaje, las cuales abarcan un amplio abanico de sectores, incluyéndose las que han sido necesarias para el desarrollo, como puede ser librerías con métodos referentes a las redes neuronales o a la limpieza de textos.

OBTENCIÓN DE PARAMETROS

En primer lugar, se comenzó el desarrollo por la parte de la preparación y descomposición de la página web a analizar.

Para este componente, lo primero que necesitamos es descargar el HTML desde la URL proporcionada por el usuario. Esto se realizará mediante la librería urllib la cual mediante una petición GET, almacenará la respuesta de dicha petición en una variable la cual contendrá la página en cuestión.

Una vez obtenida la página en formato HTML, se descompondrá ésta en distintos componentes para poder analizar. En primer lugar, se almacenará el título de la página, ya que será relevante para el análisis de la comprensibilidad del tema. En segundo lugar, se almacenarán las distintas imágenes junto con el texto asociado a la misma y la URL de la imagen.

Por último, se almacenará el cuerpo/texto de la página. Todos estos campos se almacenan en distintas matrices que contendrán los distintos datos de cada uno de los componentes de la página.

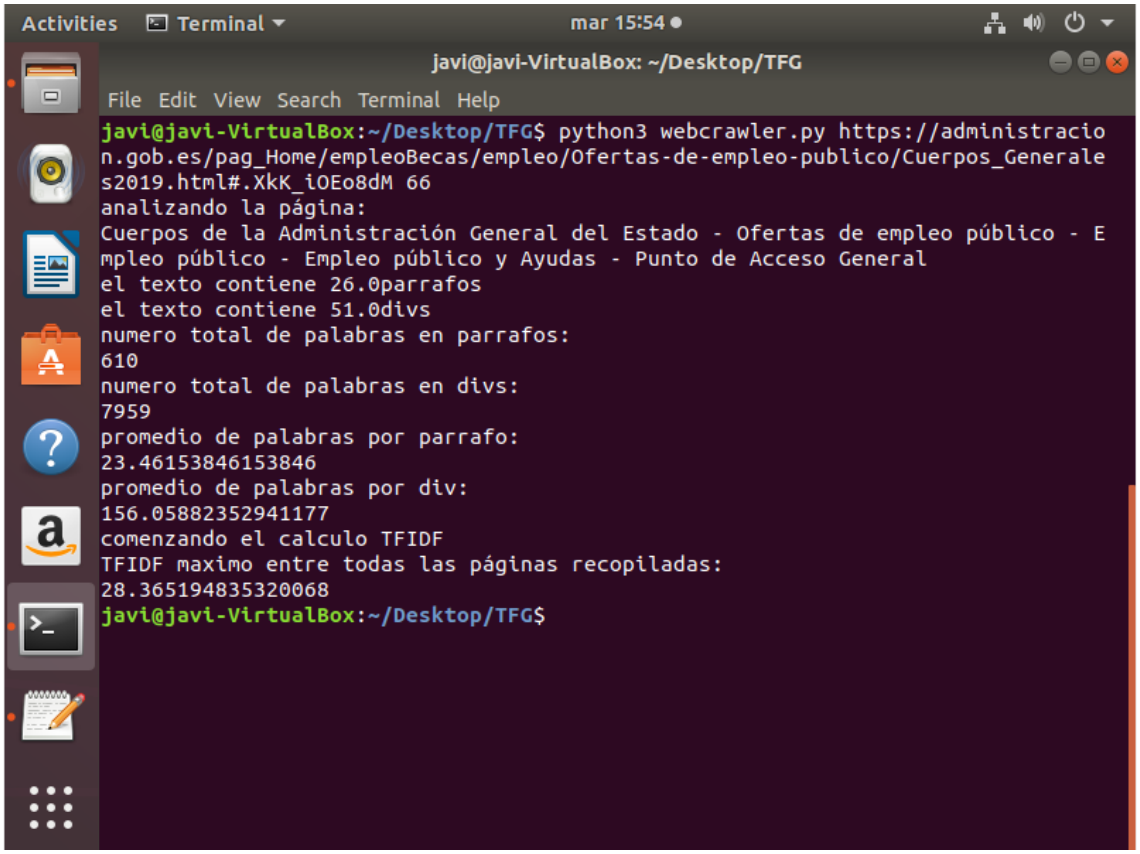


Ilustración 27Calculo parámetros terminal

Una vez obtenidos todos los campos necesarios para el análisis de la página, se procederá a la limpieza de los elementos no relevantes para el análisis, como son los elementos de puntuación o todo el código HTML (ya que los elementos del código HTML útiles han sido leídos y en función de ello, se han almacenado las variables con unos datos u otros). A su vez, palabras vacías como pueden ser los artículos se han de limpiar ya que no aportan ningún significado semántico al texto almacenada en un lista.

Toda la limpieza de texto sin significado semántico se ha realizado mediante la librería de Python BeautifulSoup, a la vez que la obtención de los distintos campos con información relevante para el análisis. Esta librería de Python es la encargada de obtener información de archivos tanto HTML como XML. Mediante dicha librería, se convierte un archivo tanto HTML como XML en un objeto de Python el cual puede ser tratado mediante dicha librería eliminando el código HTML y localizando elementos como pueden ser el título o las imágenes.

A su vez, otro de los parámetros importantes para el análisis de la comprensibilidad de los textos es el número de palabras por frase y el número de palabras por párrafo, por ello, se realizará un recuento del número de palabras encontradas en cada frase y en cada texto mediante la librería BeautifulSoup.

A continuación, se procede a la obtención de las medidas TF, IDF y al conjunto de ambas llamadas TFIDF. Esto se realiza a partir de la tokenización de las palabras del texto a analizar y la tokenización de los documentos obtenidos de los distintos portales.

Una vez tokenizados todos los textos, se procede a la obtención de los valores TF e IDF mediante las librerías de Python designadas para ello.

Una vez calculados ambos parámetros se calcula el TFIDF de cada uno de los textos de los portales con el texto a analizar.

BASE DE DATOS

A continuación, se ha procedido a la creación y preparación de la base de datos. Para la base de datos se ha utilizado una base de datos documental, concretamente MongoDB. Las bases de datos documentales se caracterizan como su propio nombre indica por almacenarse los datos en documentos. Estos documentos son conjuntos de datos semiestructurados, los cuales permiten guardar en cada documento un número de variables distinto entre documentos en la misma base de datos, y sin tener ningún tipo de dato asociado a cada variable. Esto nos permite almacenar todos los datos de una página en un solo documento y poder acceder a todos los datos obtenidos de la página en un solo documento.

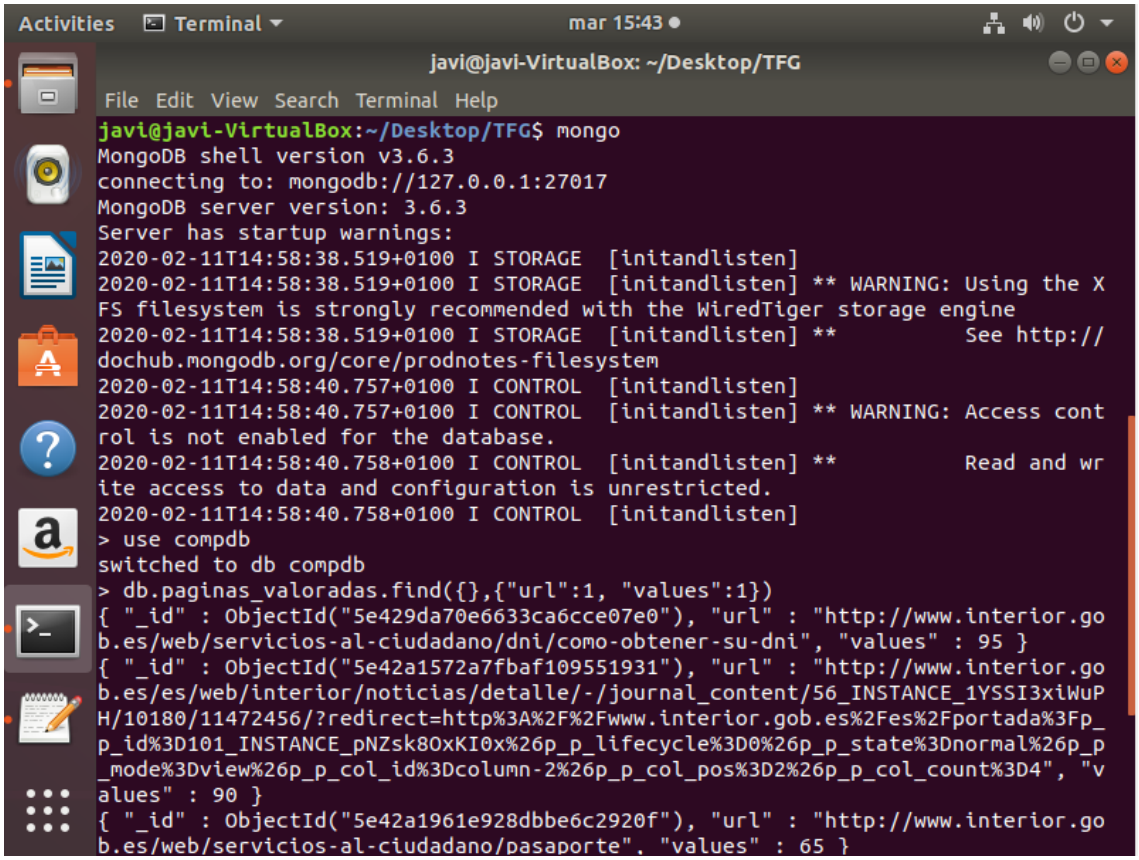


Ilustración 28Base de datos colecciones

INTELIGENCIA ARTIFICIAL

A continuación, se procederá a la creación de la inteligencia artificial. En este caso se ha seleccionado como tipo de red neuronal un perceptron. Para ello se ha utilizado la librería de sklearn.linearmodel llamada perceptron para Python. En un primer lugar, con los datos obtenidos de la base de datos (de un conjunto de datos de entrenamiento) se ha procedido al entrenamiento de la red neuronal.

El perceptron multicapa es una red de neuronas artificiales, la cual se compone de tres capas, una capa de entrada encargadas de recibir los datos o también llamados patrones, sin embargo, no realizan ningún proceso sobre los mismos.

A continuación, se encuentran las capas ocultas del perceptron multicapa, las cuales realizan un proceso no lineal sobre los patrones de entrada proporcionados por las neuronas artificiales pertenecientes a la capa de entrada.

Por último, encontramos la capa de salida, que contiene las neuronas artificiales que ofrecen el resultado final del perceptron multicapa.

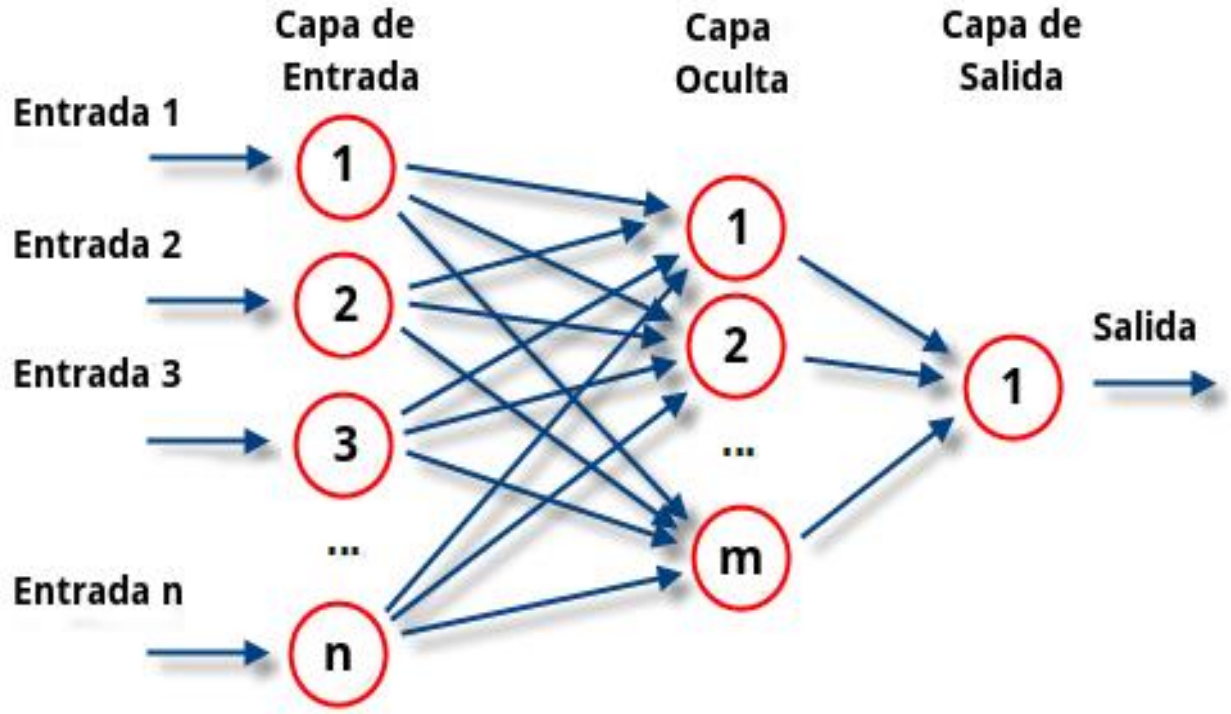


Ilustración 29 Gráfico red neuronal

En este caso, en la capa de entrada tendremos tantas neuronas artificiales como datos de entrada. El número de capas ocultas han sido calculadas en función de los distintos entrenamientos realizados sobre la inteligencia artificial. Por último, tendremos una sola neurona en la capa de salida, la cual nos devolverá la valoración final de la página web, que al ser un valor único no es necesario tener más neuronas artificiales en esta capa.

Este entrenamiento se ha realizado mediante el método *fit()* de la librería anteriormente mencionada el cual, dados unos parámetros de entrada, modifica la capa oculta de la red neuronal, ajustando dicha capa con el fin de poder realizar aproximaciones o en este caso valoraciones de los parámetros de entrada, que en este caso será la página web a analizar.

Los parámetros de entrada se obtienen de la base de datos documental creada para este sistema mediante la librería Pymongo.

```
import json
import pymongo
from sklearn.linear_model import Perceptron
import numpy as np

client = pymongo.MongoClient("mongodb://localhost:27017/")
db = client["compdb"]
pages = []
collection = db.paginas_valoradas
counter = 1
for page in collection.find():
        pages.append(page)
b = 0
X = []
y = []
Z = []
pageswovalue = []
l = 0
for b in collection.find({},{"values":0, "_id":0, "url":0}):
    pageswovalue.append(b)
fullpages = []
for page in pageswovalue:
        list_values = list(page.values())[0]
        X.append(list_values)
b=0
```

Ilustración 30 Conexión a mongoDB

Como se puede observar en este fragmento de código, en primer lugar, establecemos el cliente, al cual se le indica la URL de la base de datos. En este caso la base de datos se encuentra en el mismo servidor donde nuestro sistema está funcionando, por lo que la URL estará compuesta por *localhost*, y a su vez del puerto en el que se encuentra el proceso demonio de MongoDB, que en este caso se está utilizando el puerto por defecto el cual es el 27017.

A continuación, se selecciona la base de datos a la cual se quiere conectar. En MongoDB se pueden crear distintas bases de datos, que análogamente serían como los distintos esquemas que podemos tener en las bases de datos relacionales de Oracle.

En este caso la base de datos creada para almacenar los datos que utilizar es *compdb*.

Ahora es necesario seleccionar la colección, en este caso, el objetivo de este proceso es el entrenamiento de la red neuronal, el perceptron. Para ello accedemos a la colección paginas_valoradas, que es donde tenemos almacenados todos los parámetros de las

distintas páginas web con una valoración, tanto aportada como entrenamiento, como valorada por el propio sistema.

Mediante el método find() de la librería Pymongo, podemos hacer consultas sobre la base de datos indicada, indicándole cada uno de los parámetros que queremos o no que nos devuelva, en este caso le estamos indicando que queremos todos los valores a excepción del ID de la página, el cual es un valor obligatorio para todos los documentos introducidos en cualquier colección, que ha de ser único, la URL de la página, ya que esta no nos aporta ningún valor a la hora del cálculo de la comprensibilidad de la página y por último la valoración de la página web, ya que esta consulta se ha realizado para sacar los parámetros de entrada de la red neuronal, por lo que no se ha de introducir como parámetro de entrada la valoración dada.

Los datos devueltos por la consulta se almacenan en una lista de manera creada de forma iterativa, accediendo secuencialmente a cada una de las distintas páginas.

Una vez encontradas y almacenadas todas las páginas ya valoradas en la base de datos, esto formará nuestro conjunto de entrenamiento y de test.

Cada una de las páginas se encuentra en formato *dictionary*, es decir, contienen el nombre del campo en base de datos y el valor, en formato JSON.

```
javi@javi-VirtualBox:~/Desktop/TFG$ python3 perceptron.py
{'familiaridad': 12.0}
12.0
```

Ilustración 31Ejemplo formato JSON

En la imagen anterior, en la primera línea podemos ver el formato que devuelve la consulta sobre la página web. Este tipo de datos no puede ser insertado directamente sobre la red neuronal, por lo que es necesario un tratamiento de los datos.

Este tratamiento consiste en acceder al valor devuelto por cada uno de los parámetros y a continuación obtener solamente el valor, que se encuentra en la posición 0 del array de valores.

Mediante este tratamiento de los parámetros de entrada, obtenemos un valor, en formato *float* el cual es válido para ser introducido a la red neuronal.

Una vez obtenidos todos los parámetros de las distintas páginas web y realizado el tratamiento de cada una de las variables, se ha de formar una matriz de entrada para los datos de entrenamiento y test.

Esta matriz ha de contener en cada fila cada uno de los parámetros de cada una de las páginas web, es decir, cada fila contendrá todos los parámetros de una página web analizada.

Esta matriz ha sido creada mediante la librería Numpy, en el cual mediante la función *reshape()* la lista creada con las páginas web y las variables tratadas, se crea una matriz de las dimensiones solicitadas.

Esta matriz es la cual contendrá los parámetros de entrada de la red neuronal y ya si es posible introducirla como parámetro de entrada de la función de entrenamiento de la red neuronal.

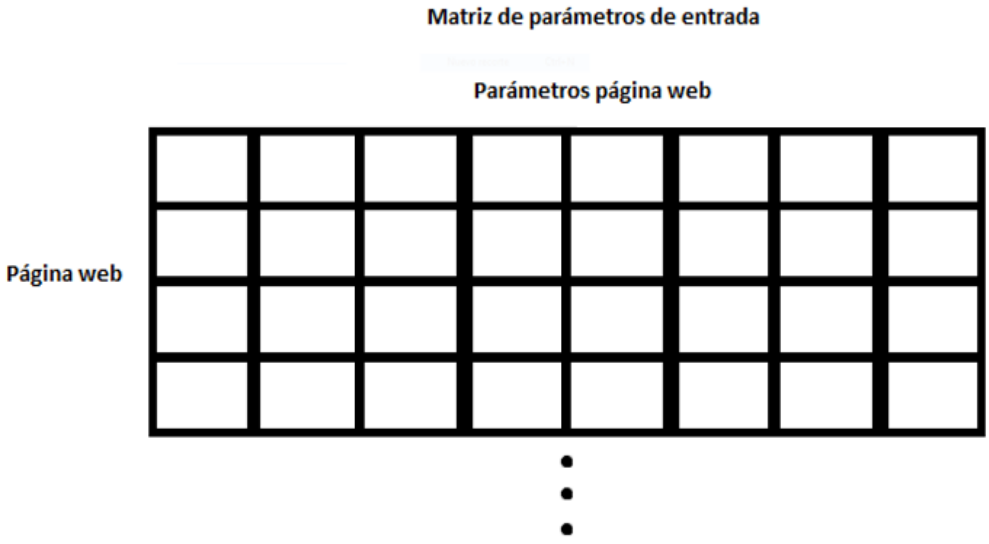


Ilustración 32Matriz entrada perceptron

Sin embargo, para el entrenamiento de la red neuronal es necesario, a su vez, introducir como parámetro de entrada de la función de entrenamiento la salida que debe dar la red neuronal, ya que el perceptron es una red neuronal que necesita de un entrenamiento supervisado para su desarrollo. Para ello, mediante la librería Pymongo, realizamos una consulta sobre todas las páginas valoradas obteniendo solamente la valoración de la página web.

Sobre esta consulta, se realizará el mismo tratamiento de los parámetros, que en este caso será un único valor para cada una de las páginas y se almacenan en una lista.

Esta lista será convertida en una matriz al igual que la matriz con los parámetros de entrada, en la cual será una matriz con una única fila, y tantas columnas como páginas en el conjunto de entrenamiento haya.

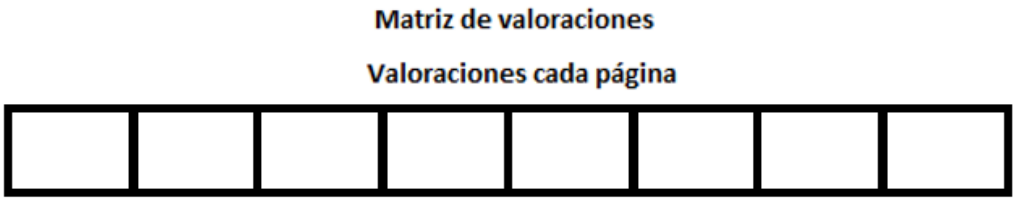


Ilustración 33Matriz salida perceptron

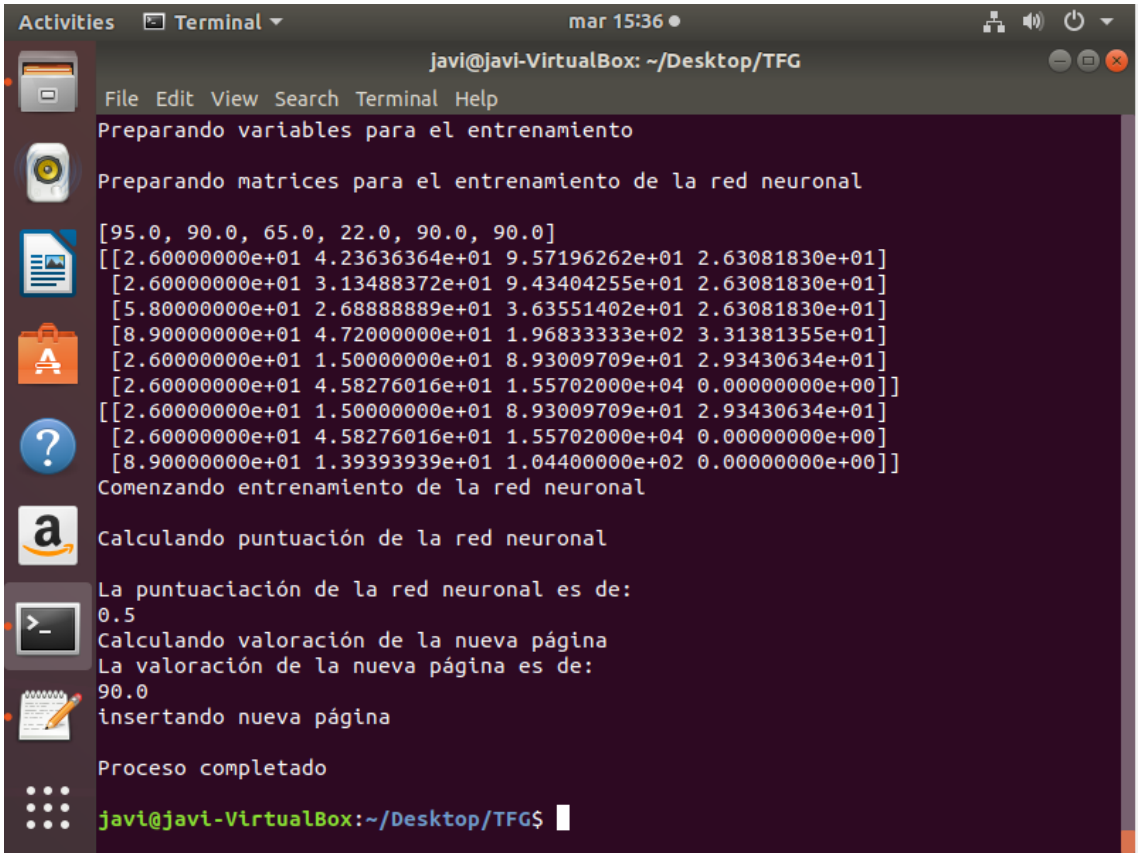


Ilustración 34Perceptron inicio

Por último, se ha diseñado la interfaz web. Esta interfaz web se ha diseñado en JavaScript. La estructura del HTML es una estructura simple en la cual se encuentran tres componentes, una casilla para introducir la página a analizar, una casilla para indicar tu edad y por último un botón para enviar la solicitud.

Una vez analizado la página, en su lugar aparecerá la misma página sin los botones y con los resultados del análisis.

Una vez se han desarrollado los distintos componentes que comprenden dicho proyecto, se ha procedido a implementar las integraciones entre los distintos componentes.

Se comenzó en un primer lugar por la integración entre el componente encargado de obtener la página a analizar y descomponerla en parámetros para introducirlos en la base de datos.

Esto se ha realizado mediante la librería pymongo de Python. El uso de dicha librería es simple ya que primero has de guardar en variables los distintos parámetros para conectar a la base de datos en cuestión que se va a utilizar. Ya que se encuentran en el mismo servidor, solo es necesario indicar a dicha librería que se ha de conectar a la máquina que se encuentra, es decir, a *localhost* seguido del puerto habilitado en MongoDB para el acceso.

Una vez indicado, se procede a indicar la base de datos en la cual se encuentra dónde queremos almacenar los datos junto con el cliente y por último los datos a almacenar, que en este caso es una sola variable en formato JSON. Mediante la librería de Python llamada json, las distintas variables obtenidas de la página se transforman a formato JSON para luego ser esta insertada en la base de datos como un documento distinto cada uno de ellos.

La siguiente integración entre componentes que se ha realizado es entre la red neuronal y la base de datos. Al haberse desarrollado la red neuronal en Python, la integración entre estos dos componentes ha sido similar a la integración entre el componente encargado de obtener la información de la página a analizar y la base de datos. Se ha utilizado la misma librería de Python llamada mongodb.

INTERFAZ GRÁFICA

La interfaz gráfica ha sido desarrollada mediante la librería Flask de Python. En primer lugar, se ha desarrollado en HTML la estructura de la página web junto con el diseño de la misma. Una vez diseñada esta estructura, se ha procedido al desarrollo de la estructura de la interfaz web. Las interfaces creadas mediante Flask, han de soportar las llamadas a cada una de las URLs pertenecientes al dominio, por lo que en el archivo *routes.py* se ha definido cada una de las rutas a las que accederá el sistema y que HTML debe cargar en cada una de las rutas, como por ejemplo en la ruta /index a la vez que en la ruta por defecto /, se ha definido que se cargue la página principal.

```
@app.route('/')
@app.route('/index')
def index():
    return render_template("index.html")
```

Ilustración 35Flask índice código

Una vez definidas todas las rutas, se han de realizar las llamadas a las funciones desde los distintos botones de la aplicación mediante los scripts creados dentro de los HTMLs mediante JavaScript.

Una vez definidos todos los parámetros, se ha de establecer la variable local de Flask para poder lanzar la interfaz y por último lanzar la aplicación.

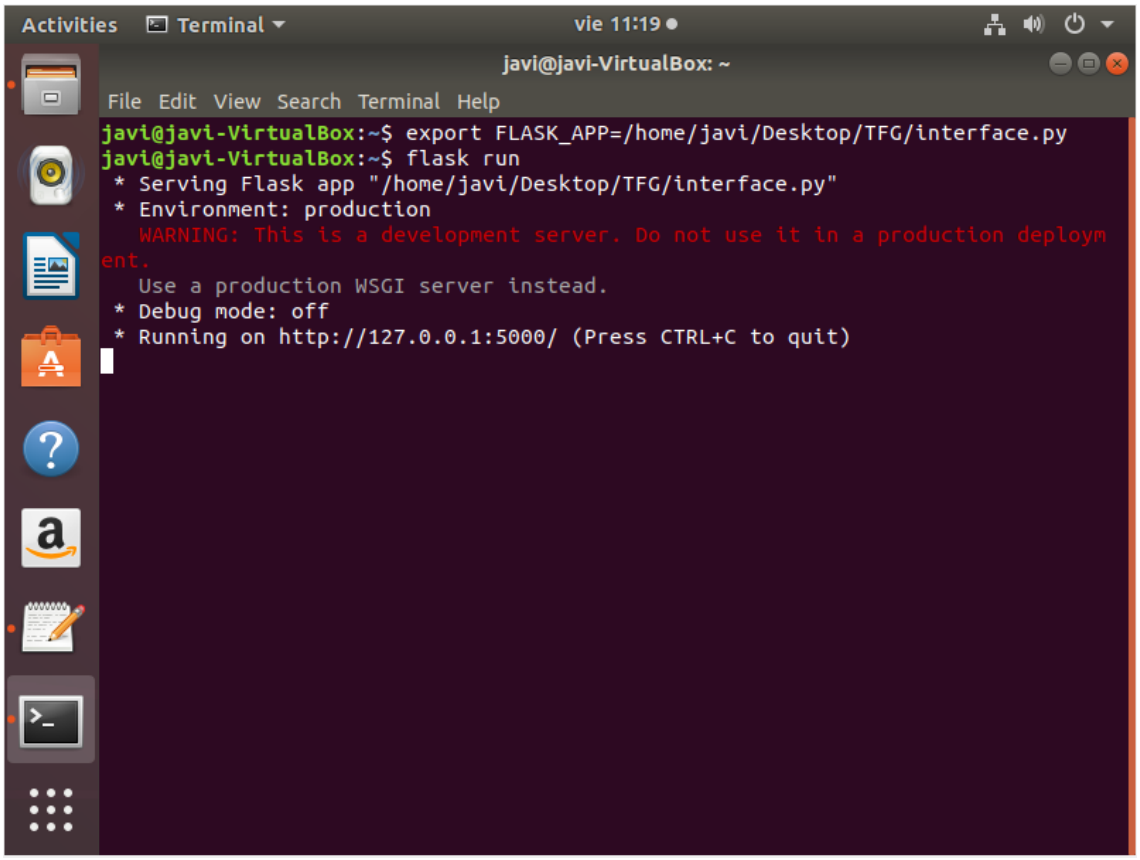


Ilustración 36Flask inicio

PLANIFICACIÓN

En este apartado se definirá la planificación del proyecto, detallando el proceso que se ha seguido para el desarrollo del sistema, junto con el presupuesto para su desarrollo.

FASES DEL PROYECTO

El Desarrollo de este Proyecto se ha llevado a cabo dividiendo el proyecto en cada una de las distintas fases que intervienen en un desarrollo.

1. En primer lugar, se ha determinado el alcance del proyecto. Se ha establecido que tipo de páginas deberá de ser capaz de analizar y el público objetivo de la herramienta.

2. Se ha dividido el proyecto en sub-tareas más pequeñas con el fin de poder planificar de manera correcta el esfuerzo que conlleva cada una de ellas. En un primer lugar, se ha dividido el proyecto en los distintos componentes que componen la herramienta, los cuales son el tratamiento de páginas web (que comprende desde el web crawler hasta la limpieza del texto), el almacenamiento de los datos (que comprende la base de datos), la red neuronal encargada de analizar los parámetros de entrada y obtener el valor de la página, la interfaz web para poder acceder a la herramienta, el spider (encargado de recolectar las páginas web de los distintos portales) y por último los conectores entre los distintos componentes y como se envía la información entre cada uno de ellos.

3. Tras la definición de los distintos componentes, se ha investigado de manera individual que tecnologías son las óptimas para cada uno de los componentes y como se van a utilizar.

4. Se han desarrollado cada uno de los componentes del sistema, obteniendo como resultado la herramienta completa. Una vez desarrollada la herramienta, se ha elaborado un plan de pruebas con el fin de cubrir todas las funcionales solicitadas por la herramienta.

5. Se ha documentado cada una de las fases descritas anteriormente en la memoria.

TABLA DE FASES DEL PROYECTO

FASE DEL PROYECTO	DEFINICIÓN DE LA FASE
Definición del alcance	En esta fase se define el alcance del proyecto entre los que destacan las páginas a analizar permitidas
Definición de los componentes de la herramienta	En esta fase se ha definido cada uno de los componentes de la herramienta y los desarrollos involucrados en cada uno de ellos
Planificación de las tareas	En esta fase se ha definido las tareas de manera más específica y se han ordenado para su realización en función de la prioridad de los requisitos asociados.
Investigación tecnologías de los componentes	En esta fase se realiza una investigación de las tecnologías óptimas para cada uno de los componentes con el fin de mejorar la eficacia del sistema
Preparación entorno de desarrollo y entorno del sistema	En esta fase se prepararán todas las herramientas necesarias para el desarrollo a la vez que se preparará el entorno donde funcionará el sistema
Desarrollo recuperación de páginas web y obtención de parámetros	En esta fase se desarrollara el web crawler encargado de recuperar la página a valorar y las distintas transformaciones de los parámetros antes de la inserción de la base de datos.

Desarrollo base de datos documental	En esta fase se instalará y se preparará la base de datos para poder almacenar los distintos parámetros necesarios para el análisis
Desarrollo interfaz web	En esta fase se desarrollará la interfaz web mediante la cual el usuario solicitará la validación y obtendrá el valor de la comprensibilidad de la página
Desarrollo inteligencia artificial	En esta fase se desarrollará la inteligencia artificial y se realizará el entrenamiento de la misma.
Desarrollo integraciones entre componentes	En esta fase se desarrollarán las integraciones entre los distintos componentes mediante los cuales estarán conectados formando el sistema final
Desarrollo de pruebas	En esta fase se desarrollaran las distintas pruebas asociadas a los requisitos
Documentación	En esta fase se realizará la creación de la documentación. Esta fase se realiza en paralelo a todas las anteriores.

Tabla 21Fases del proyecto

DIAGRAMA DE GANTT

	01/06/2019	15/06/2019	01/07/2019	15/07/2019	01/08/2019	15/08/2019	01/09/2019	15/09/2019	01/10/2019	15/10/2019	01/11/2019	15/11/2019	01/12/2019	15/12/2019	01/01/2020	15/01/2020	01/02/2020	15/02/2020
Definición del alcance	DEF ALCAN																	
Definición de los componentes de la herramienta		DEF COMP																
Planificación de las tareas			PLAN TAR															
Investigación tecnologías de los componentes				INVEST TECN														
Preparación entorno de desarrollo y entorno del sistema						PREP ENT												
Desarrollo recuperación de páginas web y obtención de parámetros							DES RECUP PAG Y PARAM											
Desarrollo base de datos documental									BASE DE DATOS									
Desarrollo interfaz web											INTER WEB							
Desarrollo inteligencia artificial												DESARROLLO INTELIGENCIA ARTIFICIAL						
Desarrollo integraciones entre componentes																INTEGRACION COMP		
Desarrollo de pruebas							PRUEBAS											
Documentación	DOCUMENTACIÓN																	

Tabla 22Diagrama de GANNT

Los sprints consistían en 15 días cada uno por ello, en el diagrama de GANTT los distintos conjuntos de fechas se han realizado mediante quincenas de días.

PRESUPUESTO

En este apartado se definirán los costes para llevar a cabo el proyecto. Estos costes serán variables en función del número de usuarios de la herramienta y de si el alcance de la herramienta aumenta.

COSTE PERSONAL

Los distintos cargos de este proyecto han sido desarrollados por la misma persona, sin embargo, se ha realizado el ejercicio de diseñar como sería el coste del proyecto con una planificación en un equipo de trabajo estándar.

El coste del personal queda definido en este apartado definiendo los costes de cada uno de los integrantes del proyecto. En siguiente tabla se contempla los costes por hora de cada uno de los distintos cargos involucrados en Euros.

CARGO	COSTE POR HORA
Administrador de sistemas	20
Jefe de proyecto	30
Ingeniero de calidad	25
Ingeniero de despliegue	20
Control de versiones	20
Desarrollador	25

Tabla 23Coste hora/cargo

En la siguiente tabla se define las horas dedicadas por cada uno de los cargos para cada una de las tareas en horas.

CARGO	INVESTIGACIÓN	DISEÑO	DESARROLLO	PRUEBAS	DOCUMENTACIÓN	DESPLIEGUE	TOTAL
Administrador de sistemas	10	25	-	-	5	5	45
Jefe de proyecto	20	15	10	5	20	5	75
Ingeniero de calidad	-	10	20	30	30	-	90
Ingeniero de despliegue	10	15	10	10	5	30	80
Control de versiones	-	5	10	15	25	-	55
Desarrollador	10	20	30	10	15	5	90
TOTAL							435

Tabla 24 Horas cargo

El coste total de cada uno de los cargos viene desglosado en la siguiente tabla junto con el coste total en Euros. Estos sueldos se han obtenido a partir de la página tusalario.es [32].

CARGO	TOTAL HORAS	COSTE HORA	COSTE TOTAL
Administrador de sistemas	45	20	900
Jefe de proyecto	75	30	2250
Ingeniero de calidad	90	25	2250
Ingeniero de despliegue	80	20	1600
Control de versiones	55	20	1100
Desarrollador	90	25	2250
TOTAL			10350

Tabla 25Coste total cargo

COSTE MATERIAL

En la siguiente tabla se puede encontrar un desglose de los costes materiales necesarios para el proyecto, junto con la previsión de gastos a un año en Euros.

MATERIAL	COSTE	COSTE EN UN AÑO
Servidor: HP Enterprise Proliant ML110 GEN10	1462	1462
Ordenador portátil: hp-pavilion-15	1199	1199
Costes indirectos	1120	12240
Material fungible	200	1200
TOTAL	3981	16101

Tabla 26Coste Material

COSTE TOTAL

En este apartado se puede observar el presupuesto total del proyecto con previsión a un año en Euros.

TIPO DE COSTE	TOTAL
COSTE DE PERSONAL	10350
COSTE DE MATERIAL	16101
PREVISION GASTOS SOBREVENIDOS	10000
TOTAL	36451

Tabla 27Coste total año 1

El coste total del proyecto ascendería a 36.451 Euros el primer año.

MARCO REGULADOR

En este apartado se realizará un análisis de los distintos aspectos legislativos que afecten al sistema o a cualquiera de los distintos componentes los cuales componen el sistema.

En el caso de este sistema, el único tipo de datos de carácter personal que se almacena es la edad del usuario, sin embargo, al no almacenarse ningún dato más del usuario, este dato no se puede asociar a ningún usuario en concreto por lo que no es posible aplicar la Ley Orgánica de Protección de Datos Personales y garantía de los derechos digitales (LOPD-GDD).

En el caso de que el sistema se haga público de manera que se utilice como web corporativa o se incluyan anuncios en la misma, se ha añadido un enlace que lleva a una página que incluye un aviso legal con el siguiente formato como se refleja en la página *ayudaleyprotecciondatos.es* [33] basado en el artículo 10 de la Ley de servicios de la Sociedad de la Información y Comercio Electrónico (LSSI-CE) [34].

AVISO LEGAL

_____, provisto con NIF/CIF _____, dirección _____, no puede asumir ninguna responsabilidad derivada del uso incorrecto, inapropiado o ilícito de la información aparecida en la página de Internet de _____

Con los límites establecidos en la ley, _____ no asume ninguna responsabilidad derivada de la falta de veracidad, integridad, actualización y precisión de los datos o informaciones que se contienen en sus páginas de Internet.

Los contenidos e información no vinculan a _____ ni constituyen opiniones, consejos o asesoramiento legal de ningún tipo pues se trata meramente de un servicio ofrecido con carácter informativo y divulgativo.

La página de Internet de _____ puede contener enlaces (links) a otras páginas de terceras partes que _____ no puede controlar. Por lo tanto, _____ no puede asumir responsabilidades por el contenido que pueda aparecer en páginas de terceros.

Los textos, imágenes, sonidos, animaciones, software y el resto de contenidos incluidos en este website son propiedad exclusiva de _____ o sus licenciantes. Cualquier acto de transmisión, distribución, cesión, reproducción, almacenamiento o comunicación pública total o parcial, debe contar con el consentimiento expreso de _____

Asimismo, para acceder a algunos de los servicios que _____ ofrece a través del website deberá proporcionar algunos datos de carácter personal. En cumplimiento de lo establecido en el Reglamento (UE) 2016/679 del Parlamento Europeo y del Consejo, de 27 de abril de 2016, relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos le informamos que, mediante la cumplimentación de los presentes formularios, sus datos personales quedarán incorporados y serán tratados en los ficheros de _____ con el fin de poderle prestar y ofrecer nuestros servicios así como para informarle de las mejoras del sitio Web. Asimismo, le informamos de la posibilidad de que ejerza los derechos de acceso, rectificación, cancelación y oposición de sus datos de carácter personal, manera gratuita mediante email a _____ o en la dirección _____

Ilustración 37Formato aviso legal [33]

IMPACTO SOCIO-ECONÓMICO

En este apartado se definirá tanto el plan de explotación del mismo, como el impacto económico, social y ético.

PLAN DE EXPLOTACIÓN

En un primer lugar, el sistema en el estado actual no genera ningún beneficio. En primer lugar, la idea es que distintos usuarios realicen solicitudes a la herramienta con el fin de que la Inteligencia Artificial del sistema siga siendo entrenada mediante las solicitudes de los usuarios y un sistema de valoración para el usuario de las valoraciones aportadas por el sistema.

Una vez obtenida una Inteligencia Artificial con una precisión muy elevada, se procederá a la inclusión de este sistema en otras herramientas como pueden ser buscadores o como parte de empresas que se dediquen al posicionamiento SEM, con el fin de poder analizar páginas de ámbito administrativo.

A su vez, al ser una herramienta orientada a procedimientos administrativos web del Estado, esta herramienta se podría incluir como proceso de validación de dichas páginas a la hora de crear una nueva página o de mantener las actuales.

Esto se realizaría estableciendo un valor mínimo de la comprensibilidad de las distintas páginas web y las que no superasen dicha valoración mínima tuviera que ser sometida a revisión, mejorando así la calidad de la comprensibilidad de las páginas web.

Esta inclusión producirá un beneficio para la gente involucrada en el desarrollo de la herramienta.

IMPACTO ECONÓMICO

Con respecto al impacto económico, esta herramienta facilitará el análisis de la comprensibilidad de las distintas páginas administrativas web del Estado, permitiendo el rediseño de dichas páginas.

Este rediseño, permitirá a los usuarios no tener la necesidad de acudir a terceros, los cuales pueden ser desde portales web que transcriben dichas páginas, hasta especialistas en procedimientos legales. El hecho de evitar que los usuarios acudan a terceros, reduciría los gastos de estos en especialistas en dichos procedimientos web.

A su vez, el cálculo de la comprensibilidad mediante una herramienta sin ánimo de lucro en el estado actual, permitirá una mayor creación de páginas administrativas web con un mayor valor con respecto a la información que aporta.

Esto provocaría una mayor agilidad en los distintos procesos administrativos, ya que el usuario entenderá los distintos procesos administrativos con mayor facilidad. El hecho de que estos procesos administrativos se realicen con una mayor agilidad, por parte del usuario evitará la posible generación de sanciones debido a que se ha excedido la fecha límite para realizar dichos trámites a la vez que evitará gastos por parte de la contratación de terceros.

A su vez, por parte del estado, la agilización de dichos procesos provocará la llegada de los pagos asociados a varios procesos administrativos web con mayor antelación, permitiendo así una reinversión de dichos beneficios en otros proyectos.

IMPACTO SOCIAL

Esta herramienta está orientada al análisis de procedimientos web del estado, por lo que la falta de inelegibilidad de los documentos de la administración digital provoca una brecha digital que impide el acceso a la misma de parte de la ciudadanía. Esto provoca guetos de población que no puede participar activamente en los servicios provistos por la web.

El hecho de poder medir la dificultad con que la información aportada por una página web orientada a procedimientos administrativos puede ser un indicador de la lentitud de muchos procedimientos administrativos, ya que, si un usuario no es capaz de poder entender los distintos procedimientos, esto impactará directamente en el tiempo de proceso de un procedimiento administrativo debido a un mal desarrollo de dichos documentos.

A su vez, al abrirse esta herramienta al público, se generarán multitud de solicitudes las cuales permitirán realizar un estudio sobre las páginas web más visitadas y a la vez que menos se entienden, ya que, si muchos usuarios solicitan la valoración de una página, esto significará que han accedido a la misma.

Por último, cabe destacar el hecho de que el análisis de las páginas web de procedimientos administrativos puede ser una herramienta útil para la valoración de páginas ya creadas o que se crearán en un futuro permitiendo solo crearse o permanecer a dichas páginas, mejorando la comprensibilidad general de todas las páginas web para los usuarios.

IMPACTO ÉTICO

Con respecto al impacto del sistema a nivel ético, el hecho de realizar una valoración sobre una página web y valorar dicha web en función de una red neuronal puede crear controversia en el hecho de que si varios grupos de audiencia no tienen discrepancia con la comprensibilidad del texto sin embargo otro grupo de audiencia podría tenerlo, el cambiar dicha página para que llegue a toda la audiencia puede hacer que el resto de los grupos de audiencia no acepten e cambio y por lo tanto la valoración no sea aceptada por todos los grupos.

CONCLUSION Y RESULTADOS

En este apartado, se realiza una conclusión de cada uno de los apartados involucrados en el desarrollo del sistema y con las opiniones personales de cada uno de los campos involucrados.

RESULTADOS OBTENIDOS

Para la valoración de la precisión de la herramienta se han realizado valoraciones sobre la comprensibilidad de diferentes páginas de procesos administrativos webs pertenecientes al Estado, las cuales no se encontraban en el conjunto de entrenamiento de la Inteligencia Artificial.

Tras el análisis de los resultados obtenidos en las pruebas de la herramienta, se ha observado que para las edades que se acercaban a edades similares a las personas encuestadas las valoraciones aportadas por la herramienta se aproximaban a las aportadas por los usuarios, sin embargo, en edades muy bajas o muy elevadas, los resultados obtenidos no se aproximaban a una valoración de la comprensibilidad real.

URL	EDAD	VALORACION SISTEMA
https://www.agenciatributaria.es/AEAT.internet/Inicio/Ayuda/_comp_Consultas_informaticas/Categorias/Presentacion_de_decla	57	77
http://www.interior.gob.es/web/servicios-al-ciudadano/asociaciones/utilidad-publica/solicitud-de-declaracion-de-utilidad-publica	16	73
http://www.interior.gob.es/web/servicios-al-ciudadano/extranjeria/ciudadanos-de-la-union-europea/numero-de-identidad-de-extra	17	90
http://www.interior.gob.es/web/servicios-al-ciudadano/asociaciones/inscripciones-registrales-de-las-asociaciones/inscripcion-de	72	88
http://www.interior.gob.es/web/servicios-al-ciudadano/indemnizaciones/responsabilidad-patrimonial-del-estado	68	36
www.interior.gob.es/web/servicios-al-ciudadano/extranjeria/control-de-fronteras/entrada-en-espana	22	39
http://www.interior.gob.es/web/servicios-al-ciudadano/dni/como-obtener-su-dni,	7	95
http://www.interior.gob.es/es/web/interior/noticias/detalle/-/journal_content/56_INSTANCE_1YSSI3xiWuPH/10180/11472456/?redir	76	90
http://www.interior.gob.es/web/servicios-al-ciudadano/pasaporte,	7	65

Ilustración 38Resultados pruebas sistema

Como se puede observar en la tabla anterior, las valoraciones de datos comprendidas entre 15 y 70 años se asemejan a valoraciones de la comprensibilidad de dichas páginas, sin embargo, valoraciones fuera de ese rango no se asemejan a valoraciones reales de los usuarios.

DESARROLLO

Principalmente el desarrollo de la herramienta ha sido lo más gratificante de todo el proyecto, ya que el poder trabajar con herramientas tan actualizadas y de tantos campos distintos, ha permitido que se pueda investigar las distintas formas de realización del proyecto, sin ninguna barrera de requisitos en el desarrollo

Se han implementado con éxito los requisitos definidos en la fase de diseño de la herramienta. A su vez, se han encontrado durante el propio desarrollo ciertos aspectos que han sido mejorados en la herramienta y otros de ellos que serán añadidos como mejoras futuras.

CALCULO DE PARÁMETROS

Uno de los principales puntos de investigación de este proyecto ha sido la definición de los parámetros influyentes en el cálculo de la facilidad de comprensión de los distintos textos. La principal dificultad de ello es mantener la objetividad de los parámetros y generalizarlos para cada uno de los rangos de edad influyentes.

Estos parámetros han sido diseñados desde el principio del desarrollo, sin embargo, durante la fase de entrenamiento y la fase de pruebas, se ha comprobado que algunos de ellos no eran influyentes sobre la comprensibilidad del texto al igual que otros que habían sido descartados desde un principio, se han añadido tras pruebas posteriores por los distintos resultados de las pruebas.

A la hora de analizar los resultados de la herramienta se ha observado que la inteligencia artificial, da un mayor peso a la edad sobre el resto de los parámetros. Se ha observado

en la fase de test que, para una misma página, analizada para distintas edades, se establece que a mayor edad mayor facilidad para entender el texto.

A su vez se ha observado que la familiaridad con las palabras del texto toma un peso bastante elevado a la hora de la inteligencia artificial realizar una evaluación.

ALMACENAMIENTO DE LOS DATOS

Con respecto a la base de datos, el almacenamiento en una base de datos documental y el trabajar con datos almacenados en la misma ha resultado más efectivo de lo esperado gracias al lenguaje de programación elegido. Se ha de destacar la importancia se seleccionar y ajustar una base de datos al tipo de datos que se va a almacenar ya que el haber trabajado con una base de datos documental ha permitido una flexibilidad en el desarrollo y una agilidad superior al mismo tiempo.

El crecimiento de las bases de datos no relacionales se comprende más tras la realización de este proyecto ya que con los motivos comentados en el párrafo anterior, las nuevas tecnologías aprovecharán más estos beneficios.

Se ha conseguido crear una base de datos que se ajuste a las necesidades del proyecto y con una eficiencia a la hora de trabajar con la base de datos que se ajusta a los requisitos solicitados.

INTELIGENCIA ARTIFICIAL

La elección de una correcta inteligencia artificial ha sido uno de los puntos más difíciles de este proyecto, ya que influyen muchos aspectos a la hora de la elección.

Desde la eficiencia de la inteligencia artificial, hasta el coste de desarrollar dicha inteligencia artificial son factores que han llevado a la elección del perceptron como red neuronal artificial a desarrollar en este proyecto.

La importancia del tiempo de respuesta de dicha red neuronal artificial ha sido un factor clave el cual, en un principio no se tuvo en cuenta a la hora de desarrollar el sistema, que finalmente ha sido crítico al querer diseñar un sistema que tuviese un tiempo de respuesta lo más pequeño posible.

Se ha conseguido entrenar una inteligencia artificial capaz de generar valoraciones que cumplen con los requisitos del proyecto, sin embargo, existen casos en los que las valoraciones de la inteligencia artificial no se ajustan a las valoraciones aportadas por los usuarios. Esto se debe a corpus de entrenamiento, ya que, con un corpus de entrenamiento mayor, la inteligencia artificial generará valoraciones con mayor precisión.

DOCUMENTACIÓN

La creación de este documento ha sido complicada, principalmente por los conceptos tan subjetivos que se han tratado en este documento. En primer lugar, obtener una definición objetiva de lo que es la comprensibilidad ha sido complicado, sin embargo, una vez obtenida dicha definición y al haber completado el desarrollo, todo se ha realizado de manera más liviana.

Hay que destacar la importancia en este aspecto de las metodologías ágiles, las cuales han permitido llevar una documentación acorde a los distintos desarrollos en cada sprint.

Esto ha agilizado el desarrollo de la documentación, haciendo que sea una tarea menos ardua y más llevadera.

PRUEBAS

Las pruebas son uno de los aspectos más importantes del desarrollo y en este proyecto se ha comprobado que todas ellas son importantes, ya que gracias a las mismas se han detectado múltiples inconsistencias entre el código y los requisitos obtenidos de cliente.

Las pruebas realizadas han comprobado con éxito los requisitos diseñados para este proyecto.

TRABAJOS FUTUROS

En esta sección se definirán las mejoras propuestas para implementar en el futuro y los distintos desarrollos que se realizarán sobre la herramienta.

ALCANCE DEL SISTEMA

En primer lugar, una de las principales ideas para implementar en este sistema es aumentar el dominio de páginas web permitidas para la valoración. En un primer lugar se incluirá la posibilidad de poder analizar paginas pertenecientes al estado, no solo a procedimientos web, sino también a páginas informativas sobre legislación o anuncios del estado sobre distintos acontecimientos.

La inclusión de dichas páginas conllevaría la inclusión de un corpus de entrenamiento con estas páginas para poder incluir dichas páginas en el entrenamiento de la inteligencia artificial.

A su vez, se debería analizar el impacto que tendría en la inteligencia artificial de estos nuevos dominios de páginas web, ya que, al haber entrenado una inteligencia artificial con solamente páginas de procedimientos del estado, podría disminuir la precisión de las valoraciones generadas para estas páginas, por lo que se debería valorar la inclusión de una segunda inteligencia artificial con el fin de analizar estos nuevos dominios de páginas web.

PRECISIÓN DE LA INTELIGENCIA ARTIFICIAL

La siguiente incorporación al sistema será el aumento del corpus de entrenamiento de la inteligencia artificial. El corpus obtenido para el entrenamiento de la inteligencia artificial no tenía un volumen muy elevado, por lo que, consiguiendo un corpus para el entrenamiento mayor, la precisión de las valoraciones generadas por la inteligencia artificial se mejoraría sustancialmente, haciendo del sistema una herramienta más fiable.

El aumento del corpus se realizará mediante crowdsourcing. El crowdsourcing consiste en delegar una serie de procesos o tareas a terceros. En el caso de la herramienta, consistiría en tras el conjunto de páginas analizadas sobre los distintos procedimientos y nuevas páginas que aparezcan sobre procedimientos administrativos, mediante crowdsourcing hacer que usuarios de distintas edades analicen la comprensibilidad de este conjunto de páginas web obteniendo así un corpus de entrenamiento más completo y con un mayor número de muestras.

Una de las principales herramientas de crowdsourcing es Figure Eight, anteriormente llamada CrowdFlower, la cual es un portal web el cual se encarga de tomar trabajos ofrecidos por distintos usuarios y empresas y hacérselo llegar a los distintos usuarios que realizarán dichas tareas.

BASE DE DATOS

Con respecto a la base de datos, una de las mejoras que se desarrollarán será la de evitar la externalización de los cálculos de los diversos parámetros de las páginas web utilizados por la inteligencia artificial con el fin de mejorar la eficiencia.

En este momento, una vez obtenidas todos los datos útiles de la página web mediante las distintas librerías de Python se calculan los diversos parámetros de la página web a analizar y estos se insertan en la base de datos.

Al utilizarse la base de datos documental MongoDB, esta permite realizar tratamiento de datos mediante diversas funciones propias de esta base de datos, permitiendo una mejora de la eficiencia en la obtención de parámetros al no externalizar dichos procesos.

ENTORNO DEL SISTEMA

Con respecto al sistema completo, este será migrado a un servidor en línea, a *Amazon Web Service* (AWS) [28] el cual nos permitirá delegar los requisitos de disponibilidad a terceros, asegurando así la disponibilidad del sistema las 24 horas del día, los 7 días de la semana.

A su vez, la capacidad de computación de estos servidores en línea nos permitirá poder procesar un mayor número de solicitudes de valoraciones de páginas web sin necesidad de aumentar el presupuesto a nivel de hardware del proyecto.

Con el aumento del uso de la herramienta en el futuro, se crearán datos sobre las solicitudes de valoraciones los cuales pueden ser analizados con el fin de la creación de estadísticas, por lo que se desarrollará un sistema de almacenamiento en base de datos, aprovechando la base de datos en MongoDB creada para el sistema con el fin de poder almacenar datos de las distintas peticiones por los usuarios con el fin de poder generar estadísticas sobre estos datos obtenidos.

PARALELIZACIÓN DE LOS PROCESOS

Otra de las optimizaciones que se incluirán en un futuro será la paralelización de los procesos ejecutados por el sistema. Al aumentar el número de hilos utilizados por el sistema y poder ejecutar diversas solicitudes de usuarios al mismo tiempo mejorara la eficiencia del sistema aumentando así el número de peticiones que se pueden procesar al mismo tiempo.

INTEGRACIONES CON OTROS SISTEMAS

Una vez implementadas estas mejoras, se investigará la posible integración de nuestro sistema con otras herramientas como pueden ser los buscadores, ya que podríamos ofrecer a estos buscadores una herramienta mediante la cual valorar la comprensibilidad de dicha página y así esta posicionarse en función de dicha comprensibilidad junto con el resto de los parámetros utilizados por estos buscadores.

GLOSARIO

URL: Dirección específica de un recurso web.

Web crawler: Programa encargado de recuperar páginas y su información de manera recursiva.

Spider: Programa encargado de recorrer páginas de manera recursiva.

Interfaz web: Conjunto de elementos que son mostrados por pantalla en un servicio web que permiten al usuario interactuar con el sistema.

Bot: Programa informático capaz de realizar tareas por sí solo.

Procesamiento del lenguaje natural: Conocimiento que trata de transformar el lenguaje de las personas para que un sistema sea capaz de entenderlo.

Scrum: Metodología ágil de desarrollo

Backlog: Lista de tareas a realizar.

Corpus: Conjunto total de documentos

Dominio: Dirección que referencia a un conjunto de páginas pertenecientes a la misma entidad.

HTML: Lenguaje de marcas para el desarrollo de páginas web

Cronjob: Funcionalidad de Linux que permite ejecutar tareas a una hora determinada de manera automática.

Tokenizar: Proceso por el cual una frase se divide en cada una de sus palabras.

Aproximador universal: Que puede detectar cualquier tipo de relación entre cualquier tipo de datos de entrada.

Inputs: Anglicismo que indica las entradas de un sistema.

Inteligencia Artificial: Sistema que presenta capacidades del ser humano.

Python: Lenguaje de programación.

Máquina virtual: Entorno dentro de un equipo que simula un sistema operativo.

Virtualización: Transformación de un sistema soportado en un entorno físico a un entorno virtual.

Linux: Sistema operativo.

Ubuntu: Distribución de Linux.

C++: Lenguaje de programación.

Bash: Lenguaje de programación basado en comandos de Linux.

Bunsenlabs: Distribución de Linux.

UML: Estándar a nivel universal para la creación de diagramas.

HTTP: Protocolo de comunicación de internet.

HTTPS: Versión del protocolo HTTP que utiliza cifrado.

NoSQL: Bases de datos no relacionales.

ACRONIMOS

UNED: Universidad Nacional de Educación a Distancia

URL: Uniform Resource Locator

HTML: HyperTex Markup Language

UML: Unified Modeling Language

PLN: Procesamiento del Lenguaje Natural

HTTP: HyperTex Transfer Protocol

HTTPS: HyperTex Transfer Protocol Secure

BIBLIOGRAFÍA

[1] V. P. M. M. S. -C. J. J.Morato, «Módulo IV Modelos de Recuperación,» UC3M, 2019. [En línea]. Available: http://ocw.uc3m.es/ingenieria-informatica/recuperacion-acceso-informacion/material-de-clase-1/MC-F-006.2.pdf.

[2] D. B. a. H. L., «web pages on children's search queries: Google vs Bing,» *Aslib Journal of Information Management*, vol. 71, nº 2, pp. 241-259, 2019.

[3] S. B. Peña, Análisis de la legibilidad lingüística de los prospectos de los medicamentos mediante el índice de Flesch-Szigriszt y la escala Inflesz, Pamplona: Anales Sis San Navarra, 2013.

[4] S. Linney, «redable.com,» 26 02 2017. [En línea]. Available: https://readable.com/blog/the-flesch-reading-ease-and-flesch-kincaid-grade-level/.

[5] A. M. Fernandez, «legible.es,» 2016. [En línea]. Available: https://legible.es/.

[6] I. Barrio, «Validación de la Escala INFLESZ para evaluar la legibilidad de los textos dirigidos a pacientes,» *An Sist Saint Navar*, vol. 31, nº 2, pp. 135-152, 2008.

[7] Semantia Labs, «www.grubric,» Madrid, 2016.

[8] R. Korntheuer, «¿Cómo medir la legibilidad de un texto?,» SEO Quito, 21 Marzo 2017. [En línea]. Available: https://seoquito.com/medir-la-legibilidad-texto/.

[9] L. Kelly, «How did the SMOG index become a crucial readability formula for content writers?,» 9 Enero 2019. [En línea]. Available: https://readable.com/blog/how-did-the-smog-index-become-a-crucial-readability-formula-for-content-writers/.

[10] M. Araque, «https://www.wearemarketing.com/,» 8 2 2017. [En línea]. Available: https://www.wearemarketing.com/es/blog/metodologia-scrum-que-es-y-como-funciona.html.

[1 K. Schwaber, «What Is Scrum?,» VOLARO, 2010. [En línea]. Available: http://www.volaroint.com/wp-content/uploads/dlm_uploads/2014/03/DC-VOLARO-Training-Scrum-What_Is_Scrum.pdf.

[1 A. Bradley, Journal of Reading Behavior, Arizona: Universidad de Arizona, 1977.

[1 P. M. T. H. Eelco Plugge, The definitive Guide to MongoDB, New York: Apress, 2010.

[1 V. P. M. S.-C. J. Jorge Morato, «Modelos de recuperación,» Universidad Carlos III, Madrid, 2019.

[1 M. Najork, «Web Crawler Architecture,» Microsoft Research, 2009. [En línea]. Available: https://www.microsoft.com/en-us/research/wp-content/uploads/2009/09/EDS-WebCrawlerArchitecture.pdf.

[1 Scrapinghub, «Scrapy,» 2015. [En línea]. Available: https://scrapy.org/.

[1 R. S.L., «Rankia.com,» [En línea]. Available: https://www.rankia.com/.

[1 A. E. d. A. Tributaria, «Agencia tributaria,» [En línea]. Available: https://www.agenciatributaria.es/.

[1 C. S.L., «comosetramita.com,» [En línea]. Available: https://comosetramita.com/.

[2 U. E. y. C. Ministerio de Asuntos Exteriores, «Ministerio de asuntos exteriores,» [En línea]. Available: http://www.exteriores.gob.es/Portal/es/Paginas/inicio.aspx.

[2 «adminfacil.es,» [En línea]. Available: https://www.adminfacil.es/.

[2 s. s. y. m. Ministerio de inclusión, «sede.seg-social.gob.es,» [En línea]. Available: https://sede.seg-social.gob.es/wps/portal/sede/sede/Ciudadanos/CiudadanoDetalle/!ut/p/z0/fYxLDoIwFACvwob1a4U0smRhGvwsjCFiN6RpH-T5aYVWo7cXOIDLmUwGFDSgnH5TryN5p-8TX5RoMy5yXjC-l6zasLI-

7OoiO2ay4nDCAFtQ_6PpQtdhUCUo413ET4QmoMV2IUfWh5TNImWGXlZb7WZDrvPjA0PyTQyOkT.

[23] N. S.L., «burbuja.info,» [En línea]. Available: https://www.burbuja.info/inmobiliaria/.

[24] D. J. Matich, «Universidad Tecnológica Nacional – Facultad Regional Rosario,» Marzo 2001. [En línea]. Available: https://www.frro.utn.edu.ar/repositorio/catedras/quimica/5_anio/orientadora1/monograias/matich-redesneuronales.pdf.

[25] R. L. B. D. M. C. B. d. M. C. Agelet de Saracibara, «Un modelo numérico para la simulación de disolución de precipitados en aleaciones de aluminio con endurecimiento utilizando redes neuronales,» *Revista Internacional de Métodos Numéricos para Cálculo y Diseño en Ingeniería*, vol. 29, nº 1, pp. 29-37, 2013.

[26] K. S. Y. O. A. R. Adil Tannouche, «A Real Time Efficient Management of Onions,» 28 Febrero 2015. [En línea]. Available: https://pdfs.semanticscholar.org/4420/e2f03be95b079897c75b4b9c35f08e906ba0.pdf.

[27] Oracle, «Virtualbox,» 2007. [En línea]. Available: https://www.virtualbox.org/.

[28] C. Ltd, «Ubuntu,» Canonical Ltd, 20 Octubre 2004. [En línea]. Available: https://ubuntu.com/.

[29] P. S. Foundation, «Python,» Python Software Foundation, 1991. [En línea]. Available: https://www.python.org/.

[30] T. B. L. Project, «bunsenlabs.org,» The BunsenLabs Linux Project, 2015. [En línea]. Available: https://www.bunsenlabs.org/.

[31] N. Project, «nltk.org,» NLTK Project, 2019. [En línea]. Available: https://www.nltk.org/.

[32] P. Jupyter, «jupyter,» Project Jupyter , 2014. [En línea]. Available: https://jupyter.org/.

[33] P. S. Foundation, «urllib — URL handling modules,» Python Software Foundation, 2001. [En línea]. Available: https://docs.python.org/3/library/urllib.html.

[34] L. Richardson, «Beautiful Soup Documentation,» 2017. [En línea]. Available: https://www.crummy.com/software/BeautifulSoup/bs4/doc/.

[35] E. L. E. K. Steven Bird, «Wordnet interface,» NLTK, 2009. [En línea]. Available: https://www.nltk.org/howto/wordnet.html.

[36] M. Inc, «MongoDB,» MongoDB Inc, 2009. [En línea]. Available: https://www.mongodb.com/es.

[37] D. Cournapeau, «scikit-learn,» INRIA, 2010. [En línea]. Available: https://scikit-learn.org/stable/.

[38] T. Oliphant, «numpy,» 1995. [En línea]. Available: https://numpy.org/.

[39] Pluralsight, «www.javascript.com,» Pluralsight, 2016. [En línea]. Available: https://www.javascript.com/.

[40] A. Ronacher, «Flask,» 2018. [En línea]. Available: https://www.palletsprojects.com/p/flask/.

[41] WagelIndicator, «tusalario.es,» WageIndicator, 2019. [En línea]. Available: https://tusalario.es/.

[42] A. P. Silverio, «ayudaleyprotecciondatos.es,» [En línea]. Available: https://ayudaleyprotecciondatos.es/modelo-aviso-legal/.

[43] A. E. B. O. d. Estado, «boe.es,» 29 12 2007. [En línea]. Available: https://www.boe.es/buscar/act.php?id=BOE-A-2002-13758&tn=1&p=20140510&vd=#a10.

[44] Amazon, «Amazon Web Service,» [En línea]. Available: https://aws.amazon.com/es/?nc2=h_lg.